The Dissertation Committee for Katherine Marie Trundt certifies that this is the approved version of the following dissertation:

CONSTRUCT BIAS IN THE DIFFERENTIAL ABILITY SCALES, SECOND EDITION (DAS-II): A COMPARISON AMONG AFRICAN AMERICAN, ASIAN, HISPANIC, AND WHITE ETHNIC GROUPS

Committee:

_____
Timothy Z. Keith, Supervisor

_____
Cindy Carlson

_____
Stephanie W. Cawthon

_____
Randy W. Kamphaus

_____
Richard R. Valencia

CONSTRUCT BIAS IN THE DIFFERENTIAL ABILITY SCALES, SECOND EDITION (DAS-II): A COMPARISON AMONG AFRICAN AMERICAN, ASIAN, HISPANIC, AND WHITE ETHNIC GROUPS

by

Katherine Marie Trundt, A.B.; B.S.; M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

August 2013

Construct Bias in the Differential Ability Scales, Second Edition (DAS-II): A comparison among African American, Asian, Hispanic, and White Ethnic Groups

Katherine Marie Trundt, Ph.D.

The University of Texas at Austin, 2013

Supervisor: Timothy Z. Keith

Intelligence testing has had a relatively long and controversial history, beginning with what is generally considered the first formal measure of intelligence, the Binet-Simon Scales (1916). Questions regarding possible cultural bias in these measures arose virtually simultaneously (e.g. Burt, 1921; Stern, 1914). Over the course of the twentieth and early twenty-first centuries, an abundance of intelligence measures have been developed, with many of them having several revisions, but the issue of test bias remains an important one, both in the professional literature and in the popular press (Reynolds & Lowe, 2009). A current intelligence measure in use, the Differential Ability Scales, Second Edition (DAS-II, Elliott, 2007), is a test with growing popularity for assessment of children and youth, not only for its ease of use, but also for its appeal to young children and its nonverbal composite (among other things). Consequently, it is essential that there be empirical evidence supporting the use of the DAS-II as an appropriate measure of cognitive abilities for children of varying backgrounds. The test publishers

conducted extensive research with a representative sample during test development in an effort to ensure that the measure met adequate reliability and validity criteria; however, the issue of test bias, particularly regarding cultural or racial/ethnic groups, was not explicitly addressed. This issue was raised and examined with the original DAS by Keith, Quirk, Schartzer, and Elliott (1999), but with the significant changes made from the first edition to the second, there is no guaranty that the evidence from the earlier would necessarily apply to the latter. The current study investigated whether the DAS-II demonstrates systematic construct bias toward children and youth of any of four ethnic groups: Black, Hispanic, Asian, and White. Multi-group confirmatory factor analysis using data from the DAS-II standardization sample was used to assess whether criteria for increasingly strict levels of invariance were met across groups. Outcomes of this research contribute to an existing body of literature on test bias, as well as provide evidence regarding cross-group construct validity in the DAS-II. Ultimately the results of this study can be used to evaluate the appropriateness of the DAS-II for clinical use with certain ethnic groups and will help to emphasize further the importance of exploring these issues with all standardized tests.

Table of Contents

Introduction

Debate and general discord have plagued the field of intelligence testing virtually since its inception. Given the complicated questions regarding what constitutes intelligence, how it can best be measured, and whether the resulting scores are meaningful and equivalent across different groups, this controversy comes as no surprise. The initiation of the psychological testing movement is traced back to Sir Francis Galton and his work in the early 1800s, though the roots of modern theories of intelligence are grounded in the work of Alfred Binet, Victor Henri, and Théodore Simon. Intelligence testing also finds its foundation in the work of these individuals, as the Binet-Simon Scale (Binet & Simon, 1905) is typically considered the first modern intelligence test (Cohen & Swerdlik, 2002).

Originally developed in France for the purposes of identifying children who would require specialized education, the Binet-Simon Scale was translated, extended, and substantially revised for use in the United States (Stanford-Binet Intelligence Scales, Terman, 1916). With these revisions and the occurrence of several coincidental events in history, intelligence tests were soon widely distributed in the United Stated and used for a variety of purposes. Despite warnings regarding the limitations of Binet's test and intelligence testing in general (Binet, 1973; Stern, 1914), these tests were often administered with little thought given to the appropriateness of their use. The limitations of these measures, in conjunction with the lack of consideration for these limitations

when administering intelligence tests to minorities, account for many of the issues and criticisms pertaining to cultural bias in intelligence testing that still exist today.

Critics of intelligence testing asserted that there was potential for language bias for minorities, due to the failure to consider the impact of English-language skills on test performance (Klineberg, 1935; Sánchez, 1934, as cited in Valencia & Suzuki, 2001, Chapter 5), in addition to broader cultural bias related to tests being oriented toward and around the majority group's experiences and values (Bond, 1987; Butler-Omololu Doster & Lahey, 1984; Thomas, 1982). Related to these arguments was an additional objection on the basis that tests measure different constructs when used with children who are not from the majority culture (Mercer, 1979). The lack of inclusion of minorities in standardization samples was also grounds for criticism, as early measures were standardized with samples of only White children (e.g., Binet-Simon, 1905; Terman, 1916; Terman & Merrill, 1937, 1960; Wechsler, 1949). Moreover, inequitable social consequences associated with the administration of intelligence tests to minorities, particularly in schools, further fueled concerns related to test bias (e.g., González, 1974, 1990; Sánchez, 1934, as cited in Valencia &Suzuki, 2001, Chapter 5; Chapman, 1988).

In spite of these fervent criticisms and the apparent decline in hereditarianism after 1930 (Garth, 1925, 1930, as cited in Valencia & Suzuki, 2001, Chapter 1; Richards, 1997), the period between 1930 and the mid-1950s was generally characterized by maintenance of the status quo regarding test bias (Valencia 1997; Valencia, 2010), perhaps in part due to the proliferation of group-administered intelligence measures (Valencia, 1997; Valencia & Aburto, 1991). The issue of test bias was relatively stagnant

until it was revived by *Brown v. Board of Education of Topeka* in 1954. Attempting to avoid desegregation as ordered through the ruling of this case, Southern schools used intelligence tests to prevent children of color from entering "white" schools (Bersoff, 1982, as cited in Valencia & Suzuki, 2001, Chapter 1). With the civil rights movement highlighting the rights of racial/ethnic minorities at approximately the same time as this attempt to use intelligence measures to circumvent desegregation, there occurred something of a "perfect storm" that helped to rekindle the test bias debate (Valencia & Suzuki, 2001, Chapter 1). Through a flurry of legislation during the late 1960s and early 1970s, followed by the initiation of related empirical research, the issue of test bias was thus reopened and carried into the contemporary era.

Two primary features of this debate were the relationships between a) curriculum differentiation (i.e., tracking) and group-administered intelligence tests, and b) overrepresentation of minority students in special education (Valencia, 1999; 2008). These concerns, combined with the previously mentioned objections and questions surrounding the interpretation of differences in group performance on intelligence measures (e.g., Eells, Davis, Havighurst, Herrick, & Taylor, 1951) laid the groundwork for an entire body of research investigating test bias (for a review see e.g., Jensen, 1980; Reynolds & Lowe, 2009; Valencia & Suzuki, 2001, Chapter 5). As technological and methodological advances were made, the conceptualization of test bias, particularly cultural bias, moved in the direction of differential psychometric or statistical validity, and away from more subjective questions of cultural loading and test fairness. From this

new perspective, test bias could be investigated psychometrically in terms of content validity, construct validity, and predictive validity.

Before proceeding it is important to note that although validity and test bias are obviously connected, a significant distinction to make between the two is that bias involves comparison between two (or more) groups, while validity can apply to only one group (Jensen, 1980). This observation is especially relevant when considering the "test bias" definitions of each type of validity, as empirical test bias can apply to any type of group (i.e. groups based on race/ethnicity, sex, socioeconomic status, etc.). Generally, however, when one is comparing ethnic or racial groups test bias is referred to as cultural bias (e.g., Valencia & Suzuki, 2001, Chapter 5). Furthermore, when comparing ethnic or racial groups, the comparison is typically between two groups, often with one identified as the "major" group and the other labeled as the "minor" group. Jensen (1980) clarified these distinctions, as they are not indicative of value judgments:

> The major group can usually be thought of as (1) the larger of the two groups in the total population, (2) the group on which the test was primarily standardized, or (3) the group with the higher mean score on the test, assuming the major and minor groups differ in means (p. 376).

Valencia and Suzuki (2001, Chapter 5) added an additional distinction: "The major group is the group that the test is believed not to be biased against" (p. 117).

Although the measures discussed thus far are now largely outdated, contemporary measures of intelligence comprise the majority of the research conducted regarding cultural bias. A number of reviews exist regarding this research, with many providing

evidence counter to many of the criticisms presented previously and concluding that evidence generally supports a lack of cultural bias against American, English-speaking racial/ethnic minority groups on cognitive measures and (e.g., Brown, Reynolds, & Whitaker, 1999; Jensen, 1980). These findings, however, are not without some limitations. Valencia and Suzuki (2001, Chapter 5) have presented a detailed review of this research, and for the sake of establishing some degree of background and gaining a general sense of the status of the literature and limitations within this research, a brief summary of their conclusions is provided here.

In their conclusions, Valencia and Suzuki observed that identification of cultural bias appeared to be "psychometric specific," in that the frequency of findings of bias (and, conversely, findings of non-bias) tended to coincide with certain types of validity and not others. Although it is difficult to draw any conclusive inferences from this observation, the distribution of findings could potentially be an artifact of the methodologies used to study bias, the definitions themselves, or some other factor. Valencia and Suzuki also recognized a number of limitations to the current cultural bias literature, criticizing the lack of research with minority groups other than Black and Hispanic, and the lack of geographical representation in cultural bias research (only 12 of the 50 states were represented). They suggested that there may be problems with the external validity of existing research, as most of the investigations into cultural bias were conducted with children in general education, despite the fact that most of the tests are used to evaluate children who are referred for special education. Moreover, whereas most studies examined bias across racial or ethnic groups, very few controlled for or

addressed potential confounding issues of socioeconomic status, language dominance, and/or sex. Valencia and Suzuki's final point offered caveats for the "sweeping claims" of a general lack of cultural bias overall made by Jensen (1980). They acknowledged that there is a lot of evidence supporting a lack of cultural bias in intelligence tests, but in light of the weaknesses in the literature, they asserted that broad claims of completely unbiased measures are premature and the possibility for cultural bias in intelligence tests remains an open issue (Valencia & Suzuki, 2001, Chapter 5; also, Jensen, 1980; Reynolds & Low, 2009).

An additional observation that Valencia and Suzuki made is that in general, research into test bias appears to be on the decline, particularly when one compares the number of studies conducted in the 1970s and 1980s to the number conducted in the 1990s forward. This waning of research could be due, in part, to the "sweeping inferences" of non-bias presented in Jensen's well-known work (1980). Additional presumptions of non-bias may also stem from more recent efforts often made when norming a cognitive test to ensure adequate representation within the norming sample, as well as in-depth analyses into various aspects of reliability and validity during development and standardization. Regardless, despite the recent decline in the popularity of this type of research, the importance to "press on" has not waned (Reynolds & Lowe, 2009; Suzuki & Valencia, 1997).

Valencia and Suzuki identified investigations into cultural bias for fourteen different intelligence measures, many of which have since been revised once (e.g. Elliott, 2007; Kaufman & Kaufman, 2004) or in some cases twice (e.g., Wechsler, 1991, 2004;

Woodcock & Johnson, 1989; Woodcock, McGrew, & Mather, 2001). Most of these updated and current measures have not been evaluated for the presence of cultural bias, despite the likelihood of being administered with increasing frequency to diverse populations as globalization redefines boundaries of the world and the demographics within the United States evolve. The importance of evaluating cognitive measures for cultural bias cannot be emphasized enough. These intelligence tests are among the most popular measures that psychologists administer (Stinnett, Havey, & Oehler-Stinnett, 1994; Wilson & Reschly, 1996) and are used for purposes as diverse as determining eligibility for special education services, identification of an individual's need for services, determination of parameters (i.e. intensity and duration) of treatment (Dowdy, Mays, Kamphaus, & Reynolds, 2009), political advocacy, program evaluation, and research (Keough, 1994). Even the Wechsler scales, as the measures with the most test bias research, have more evidence investigating bias than any other measure, but this evidence is far from complete (Valencia & Suzuki, 2001, Chapter 5).

A contemporary measure currently in use, the Differential Ability Scales, Second Edition (DAS-II, Elliott, 2007), is a relatively recent revision of the original Differential Ability Scales (DAS; Elliott, 1990) and is also an indirect descendant of both editions of the British Ability Scales (BAS; Elliott, Murray, & Pearson, 1979; BAS-II; Elliott, 1996). As a test with growing popularity, not only for its ease of use, but also for its appeal to young children and its nonverbal composite (among other things), it is essential that there be evidence supporting the use of the DAS-II as an appropriate measure of cognitive abilities for children of varying backgrounds. In an effort to ensure that the measure met

adequate reliability and validity criteria, the test publishers conducted extensive research with an over-representative sample of minority children during test development. By including a greater proportion of minority participants in their sample than that found in the general population, the authors attempted to ensure adequate representation of these groups for statistical purposes; however, the issue of cultural test bias was not explicitly addressed. There is some evidence for construct validity in the DAS-II (Elliott, 2007; Keith, Low, Reynolds, Patel, & Ridley, 2010; Keith, Reynolds, Roberts, Winter, & Austin, 2011), but there has not been any research to explore the possibility of cultural bias (i.e., determining whether the construct validity holds across cultural groups). This issue was raised with the original DAS and addressed by Keith, Quirk, Schartzer, and Elliott (1999), but, again, it has not been examined with the newer edition.

In order to address this issue, the current study investigated whether the DAS-II demonstrates systematic construct bias toward children of any of four ethnic groups: Black, Hispanic, White, and Asian. Multi-group confirmatory factor analysis using data from the DAS-II standardization sample was used to assess whether criteria for increasingly strict levels of invariance are met across groups. These analyses were used to determine whether the DAS-II measures the same constructs across groups, and thus tests for construct bias across groups. Results of this research will contribute to an existing body of literature on test bias, as well as provide evidence for or against the presence of cross-group construct validity in the DAS-II. Ultimately the findings of this study can be used to evaluate the appropriateness of the DAS-II for clinical use with certain ethnic

groups and will help to emphasize further the importance of exploring these issues with all standardized tests.

Method

Instrumentation

The Differential Ability Scales, Second Edition (DAS-II, Elliott, 2007) is an individually administered test of cognitive abilities for children and adolescents ages two years, six months (2:6) to 17:11. The test is comprised of twenty-one subtests divided into two overlapping batteries: the Early Years Battery for children ages 2:6 through 6:11, and the School-Age Battery for children ages 7:0 through 17:11. Within the Early Years Battery is an additional level of differentiation, with a lower level for very young children (2:6-3:5) and an upper level for slightly older children (3:6-6:11). All batteries yield an overall composite score, lower-level diagnostic "cluster" scores, and specific ability measures, which include both the core and diagnostic subtests.

The DAS-II was revised from its original version in order to provide updated norms that were achieved using a sample that was representative of the current population, as well as "to address more fully some of the current trends in cognitive and developmental theory" (Elliott, 2007, p. 1). In addition to the use of a new norming sample, notable changes from the original DAS include four new subtests: Rapid Naming and Phonological Processing reflect developments in research relating to dyslexia, while Recall of Digits Backward and Recall of Sequential Order reflect more current research in the area of working memory. Intended for use with exceptional children (i.e., children with attention difficulties, mild verbal or language limitations, etc.) the DAS-II provides measures of both general nonverbal conceptual and reasoning abilities and can be used to

compare ability measures to a wide range of academic achievement measures. A description of each subtest, based in part on information from the DAS-II technical manual (Elliott, 2007), is provided in Table 1.

Table 1

Description of the DAS-II Subtests

Subtest	Description
Verbal Comprehension	Measure of receptive language for which child follows oral instructions to point to, manipulate, and select objects or pictures
Naming Vocabulary	Expressive language measure for which child names objects presented in pictures
Word Definitions*	Measure of knowledge of word meanings for which the child explains the meaning of words presented orally by the examiner
Verbal Similarities*	Measure of word knowledge and verbal reasoning skills on which child explains how three named things or concepts go together
Early Number Concepts	Child answers basic quantitative questions (counting, number concepts, and arithmetic) presented orally with corresponding pictures
Picture Similarities	Nonverbal reasoning measure for which child matches pictures based on concrete and abstract relationships by placing response card below one of four pictured stimuli
Matrices*	Measure of nonverbal reasoning for which child solves visual puzzles by selecting image missing from a 2x2 or 3x3 matrix
Sequential and Quantitative Reasoning*	Measure of ability to detect sequential patterns for which child determines which image completes a sequence of pictures, numbers, or geometric figures
Copying	Measure of visual perception and motor coordination for which child draws a reproduction of abstract, geometric line images
Matching Letter-Like Forms	Measure of visual perception and discrimination requiring child to match shape of presented figure to identical shape presented as part of a series of similar options

Table 1 (cont.)

Pattern Construction*	Measure of visuo-perceptual matching, nonverbal reasoning, and spatial visualization for which child uses wooden blocks, plastic blocks, or flat tiles to recreate constructions made by examiner or to recreate patterns presented in pictures
Recall of Designs*	Measure of visuo-motor integration and short-term recall fro which child reproduces drawing of abstract geographic designs from memory after viewing drawing for 5 seconds
Recognition of Pictures*	Measure of short-term, nonverbal visual memory for which child views images for a set period of time before being asked to select images viewed from a set of distracters
Recall of Objects-Immediate*	Measure of short-term memory for which child is taught names of 20 pictures immediately before being given 40-45 seconds to recall as many pictures as possible
Recall of Objects-Delayed*	Measure of immediate-term memory for which, without viewing pictures again, child is asked to recall pictures after a 10-30 minute delay
Recall- Digits Forward*	Measure of short-term, auditory memory for which child repeats increasingly long series of digits recited by examiner
Recall- Digits Backwards*	Measure of short-term, auditory memory for which, in reverse order, child repeats increasingly long series of digits recited by examiner
Recall of Sequential Order*	Measure of short-term recall for which child recalls, in a specified sequential order, increasingly long series of verbal and pictorial information
Speed of Information Processing*	Measure of efficiency of performing simple mental operations for which child marks figure with most parts in each row as quickly as possible
Rapid Naming*	Measure of efficiency of naming visual stimuli for which child names colors and/or images as quickly as possible without making mistakes
Phonological Processing	Measure of awareness of and ability to manipulate phonemes in the English language for which child responds to verbal stimuli by rhyming, blending sounds, deleting sounds, and identifying phonemes in words

Note. *Denotes subtests included in the current analyses

Development of the instrument. According to the DAS-II manual (Elliott, 2007), development of the DAS-II was an iterative process, beginning with focus groups and surveys of experts and examiners during conceptual development, followed by a national pilot and a national tryout, standardization, and ultimately, scaling and norms development. During the conceptual development phase, special emphasis was placed on identifying the primary goals of revision, which were determined to be, "to update the normative information and kit materials while maintaining the integrity and overall design of the instrument" (Elliott, 2007, p 101). Later, during the national pilot and tryout phases, additional emphasis was placed on considering evidence of clinical utility, as well as bias at the item level. Results from the national pilot and national tryout were used to inform the development of the standardization edition of the measure. The Rasch model of item response theory was employed to analyze items and subtests and to develop item-scoring rules. Moreover, data supporting reliability, validity, and clinical utility of the measure were also collected during the standardization phase.

Reliability and validity of the DAS-II. In general, evidence provided in the testing manual suggests that the DAS-II demonstrates at least adequate reliability, and it frequently meets standards of excellence (Elliott, 2007). For the overall sample, the range of average reliability (internal consistency) coefficients for subtests on the Early Years battery was from .79 to .94. For the School Age battery, average reliability coefficients ranged from .74 to .96. Overall, these reliability estimates are improved from the original edition of the measure. Across age groups, average corrected stability coefficients of subtest, cluster, and composite scores range from .63 (two subtests,

Matching Letter-Like Forms and Recognition of Pictures) to .91 (one subtest, Naming Vocabulary, and the General Conceptual Ability) over a retest interval of 1 to 9 weeks. All cluster and composite scores have corrected stability coefficients between .81 and .92, which are considered to be in the "excellent" range.

As is true for reliability, the DAS-II also demonstrated strong evidence of validity (Elliott, 2007). To begin, the Cattell-Horn-Carroll theory—the model of intelligence that informed the development of the DAS-II and the model used for the current study—is the model of intelligence with the most empirical support and although imperfect, is "the best current description of the structure of human intelligence" (Keith & Reynolds, 2010, p 8). With that in mind, for the overall sample and across age groups, evidence of (internal) convergent and discriminant validity was provided by demonstrating the expected pattern of correlations among subtests and composites. Specifically, higher correlations were found among subtests comprising the same composite (e.g., Verbal Ability) than with subtests comprising other composites (e.g., Verbal with Spatial Ability). Confirmatory factor analysis was also conducted in order to establish evidence for the test's factor structure. Evidence supporting a two-factor hierarchical model was established for children ages 2:6-3:5, a four-factor model for children ages 4:0-5:11, and a six-factor model for children ages 6:0-17:11. These analyses were broken down according to the specified age groups due to differences in the subtests that were administered to children of different ages. Independent researchers (Keith et. al., 2010) have more recently used additional factor analyses to demonstrate strong evidence for a six-factor model for children ages 4:0-17:11.

External validity (i.e., concurrent validity) was assessed by comparing DAS-II scores with scores from other ability measures (Elliott, 2007). As described in the test manual, several expectations were met with regard to these comparisons. First, the scores of the DAS-II were moderately to strongly correlated with scores from other comprehensive measures of cognitive abilities, ranging from .59 with the Bayley Scales of Infant and Toddler Development, Third Edition (Bayley-III; Bayley, 2005) to .84 with the Wechsler Intelligence Scale for Children, Fourth Edition (WISC-IV; Wechsler, 2003) to .88 with the original DAS. Second, The DAS-II correlated moderately with total scores from various measures of academic achievement, ranging from .79 to .82 with the Wechsler Individual Achievement Test, Second Edition (WIAT-II; Wechsler, 2001) for a sample of children identified as having ADHD and LD and a non-clinical sample, respectively. Additional evidence supporting the validity of the DAS-II was provided through independent research conducted to examine sex differences on the DAS-II (Keith, Reynolds, Roberts, Winter, & Austin, 2011). The authors investigated construct bias by testing for measurement invariance as a preliminary step, with sufficient levels of invariance established across groups.

Participants

Participants in the current study were cases selected from the DAS-II standardization sample. The overall standardization sample was stratified according to age, sex, race/ethnicity, parent education level, and geographic region based on data gathered in 2005 by the US Bureau of the Census (Current Population Survey), with oversampling of minority ethnic groups. The sample from which cases were drawn for

the current study included 2,270 children ages 5:0-17:11 from Hispanic, Black, Asian, and White ethnic groups.[1] Sample sizes for each ethnic group are provided in Table 2. Because there were many more White children in the total sample, White participants were selected at random from the total sample to form two White sub-samples to be equal in size to the largest of the other sample sizes.

Details regarding countries of origin or countries of heritage for minority children, particularly those of Asian and Hispanic descent, were not available. According to the test manual, all children spoke English, with bilingual children included only if English was reported to be the primary language of the child; however, the test manual did not report details regarding how a child's primary language was determined. Additionally, all children were able to communicate verbally at a level consistent with their age; no nonverbal or uncommunicative children were included.

Table 2

Demographic Characteristics of Study Participants

	Variable	Asian	Black	Hispanic	White 1	White 2	Total
	N	98	407	432	432	432	1,801
Sex	Male	39	198	218	228	213	896
		(39.8%)	(48.6%)	(50.5%)	(52.8%)	(49.3%)	(49.8%)
	Female	59	209	214	204	219	905
		(60.2%)	(51.4%)	(49.5%)	(47.2%)	(50.7%)	(50.2%)
Age	5:0-7:11	26	91	114	91	76	398
		(26.5%)	(22.4%)	(26.4%)	(21.0%)	(17.6%)	(22.1%)
	8:0-10:11	20	95	104	94	100	413
		(20.4%)	(23.4%)	(24.1%)	(21.7%)	(23.2%)	(22.9%)
	11:0-13:11	21	95	98	94	112	420
		(21.4%)	(23.9%)	(22.6%)	(21.8%)	(26.0%)	(23.3%)
	14:0-17:11	31	126	116	153	144	570
		(31.7%)	(31.1%)	(26.8%)	(35.5%)	(33.3%)	(31.6%)

[1] Group labels used are the same as those used in the DAS-II technical manual (Elliott, 2007).

Procedure

The current study used multi-group confirmatory factor analysis (MG-CFA) to investigate whether the DAS-II demonstrates construct bias toward children of any of four ethnic groups: Asian, Black, Hispanic, and White. Assessing for construct bias—also called measurement bias—across groups investigates whether the test measures the same underlying construct in the same way for all groups under consideration. This is equivalent to testing for measurement invariance across groups—evaluating whether the measure varies in the underlying construct it measures based on group membership. Measurement invariance is most frequently evaluated using MG-CFA (e.g., Cheung & Rensvold, 2002), which "provides a more organized, direct, objective, and complete method for detecting construct bias than do other methods" (Keith, et. al., 1995, p. 347). See Appendix for a review of alternative methods for evaluating test bias.

Analysis

Descriptive statistics. Correlation matrices, means, and standard deviations of subtest standard scores were calculated for each ethnic group using IBM SPSS Statistics, Version 19.0 (IBM Corp., 2010). Covariance matrices were analyzed via SPSS Amos, Version 19.0 (Arbuckle, 2006) using multi-group confirmatory factor analysis. Missing data was very minimal, with a total of only 8 missing individual subtest standard scores in the total sample of 1,801 participants (each with 13 individual subtest standard scores). In order to obtain modification indices through Amos, however, there cannot be any missing data. Thus, full-information maximum likelihood was used to obtain means,

variances and covariances from the raw data in order to create covariance matrices that were then used to conduct the analyses in Amos (e.g., Graham & Coffman, 2012).

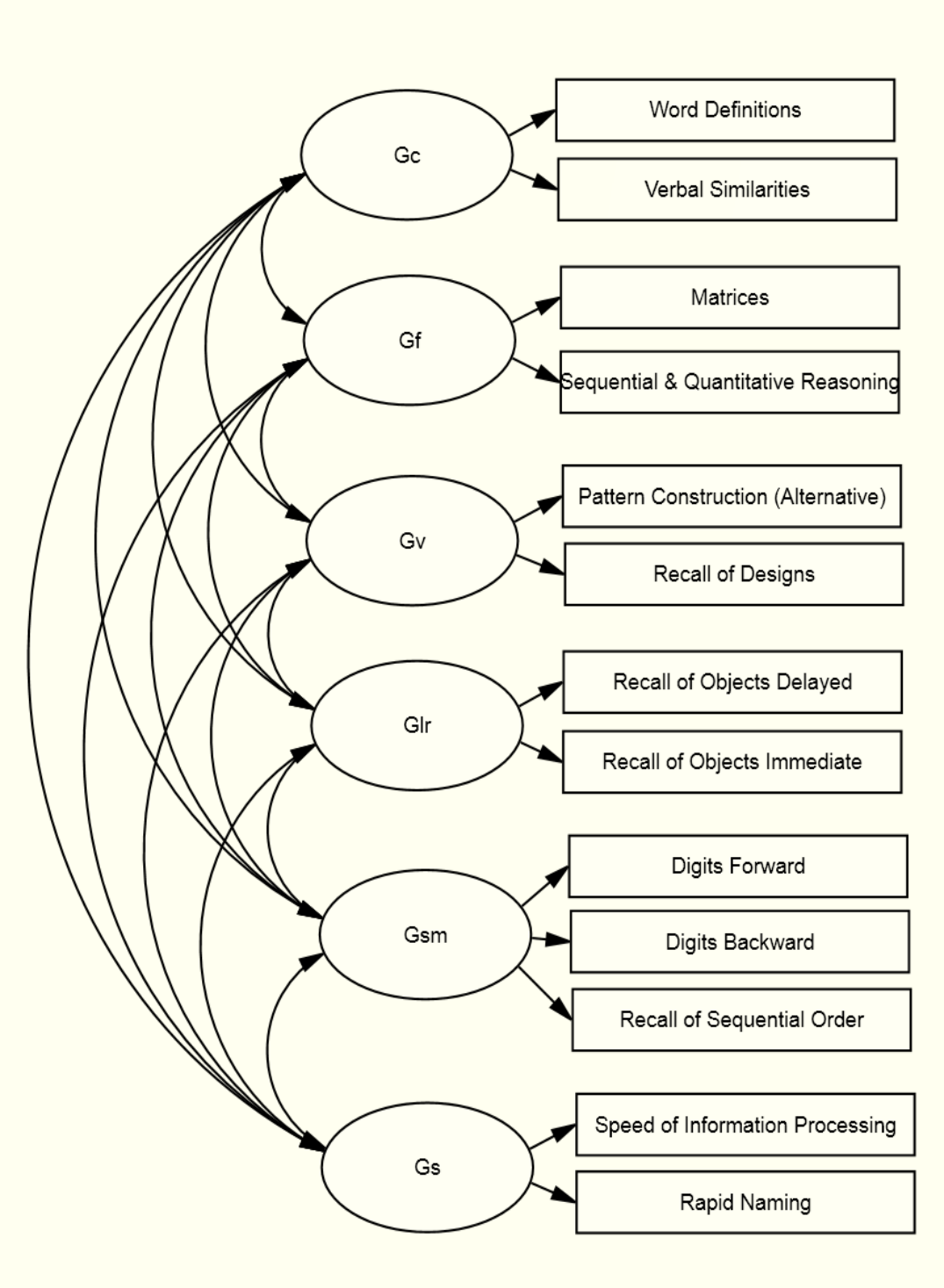
Data analysis. The purpose of this research was to determine whether the DAS-II measures the same constructs for Black, Hispanic, Asian, and White children ages 5:0 to 17:11. In other words, by testing for construct bias or measurement invariance across racial/ethnic groups, the current study evaluated whether there is evidence that the DAS-II is biased against any of these groups. The model used for these analyses is illustrated in Figure 1, with subtests limited to those administered to all children within specified age range. To clarify, the DAS-II includes some subtests for school-age children that are only administered to certain age subsets within the 5:0-17:11 range (i.e., school readiness measures for younger children), and thus subtests only administered to portions of the sample were not included in the current analyses. As mentioned previously, multi-group confirmatory factor analysis was used to address this question and did so by testing increasingly strict levels of invariance across ethnic groups. Group comparisons were conducted in a pairwise fashion, so as to compare the majority group (White) individually with each minority group (Black, Hispanic, and Asian), as is common in this type of research (e.g., Jensen, 1980). Procedures were replicated with an additional White comparison sample in order to explore the reliability of initial findings.

Each level of invariance was progressively stricter than the previous level, meaning that with each step of invariance testing additional equality constraints were imposed across groups. Criteria for lesser degrees of invariance must be met—in other words, invariance across groups given fewer equality constraints must be established—

prior to advancing to the next level of analysis. Generally following the process described by Meredith (1993), specific levels of invariance, in order of increasingly stringent criteria, include configural invariance, metric invariance, and intercept invariance. If invariance is not present across groups at any level, and partial invariance cannot be established, then the test is considered biased toward that particular group. Due to the increasingly stringent criteria required to progress through each step of these analyses, the current study addressed two specific research questions: (a) Is construct bias present in the DAS-II? And (b) If so, where does this bias exist? Additionally, given the pairwise nature of the comparisons and the need to decrease the White sample size to match that of the largest ethnic minority sample, a third question was addressed: (c) Can these findings be supported via replication with a different comparison group? The following sections describe each level of invariance testing in more detail and clarify how these analyses addressed the above research questions.

Configural invariance. Generally speaking, to establish configural invariance is to establish that the measure's factors have the same configuration across groups. More specifically, analysis at this level assessed whether the constructs that a test measures have the same configuration, or pattern of factor loadings, across groups. To do this, the factors and the pattern of loadings were set to be the same for all groups, but no equality constraints were made. In other words, the model structure was the same for both groups, but the parameters estimated across models (means, intercepts, variances, etc.) were free to vary. Means of all latent variables were fixed to zero, while the intercepts for the

Figure 1. Factor Structure of the DAS-II



Note. Error terms not shown.

measured variables were freely estimated for all groups. In order to set a scale for each factor, one unstandardized factor loading for each factor was set to one, while the remaining unstandardized factor loadings were allowed to vary freely. This method, called unit loading identification, was used throughout the current study because the alternative method for standardizing a factor— unit variance identification, which fixes the factor variance to 1.0— is most appropriate for use with factors in standardized form. For the purposes of the proposed study, analyses were conducted on factors in unstandardized form (for reasons which will be made clear later), making unit loading identification the more appropriate method. Measures of overall model fit were considered before moving to the next step of the analyses. An explanation of model fit, the various indicators of model fit, and the criteria used to determine good model fit are explained in the section following the discussion of each level of invariance testing.

Metric invariance. Only after meeting criteria for configural invariance was metric invariance (or *weak factorial invariance*) assessed. This level of invariance testing examines whether the relation of the measured variables to the latent variables, or the scale of the latent variables, is the same across groups. To do this, the unstandardized factor loadings—which were estimated freely in the Configural model—were constrained to be equal to corresponding factor loadings across groups. As in the first level of invariance testing, measures of overall model fit were again considered, although at this level comparisons of model fit were also made. Details regarding these comparisons are described in the section following intercept invariance, but in short, a

significant decline in model fit from the Configural model to the Metric model suggests a lack of support for metric invariance. As Keith and Reynolds (2012) explained, if the additional constraint in the Metric model is supported (i.e., model fit does not significantly worsen) then a one unit increase in a specific latent variable should result in an increase on each associated measured variable that is the same across groups. If acceptable model fit cannot be achieved at the metric invariance level, partial invariance can be explored; however, this procedure will be described in more depth later. Assuming support for the presence of metric invariance, one can then conclude that the underlying factors are the "same" across groups (Keith & Reynolds, 2012). However, assessing intercept invariance is also essential, as metric invariance does not provide sufficient evidence to support claims of a lack of construct bias. More specifically, one cannot claim (based on metric invariance) that all groups with the same levels of latent abilities would have the same observed scores (Keith & Reynolds, 2012).

Intercept invariance. After criteria for metric invariance was met, intercept invariance (or *strong factorial invariance)*, was assessed. Invariance at this level of analysis suggests that, given equal scores on latent factors, the intercepts of the measured variables are the same across groups. As described in Keith and Reynolds (2012), differences in intercepts—and therefore means—on measured variables are indicative of a systematic advantage for one group over another, as one group would display higher means for subtests than another group despite having equal latent abilities across groups. Previously, in the Metric invariance model, factor means were fixed to zero (and thus were set to be equal across groups), and subtest intercepts were allowed to vary freely

across groups. To make the assessment associated with intercept invariance, however, factor means were allowed to vary across groups and all corresponding subtest intercepts are constrained to be equal across groups. As Keith and Reynolds (2012) indicated, differences in latent (factor) means should account for the differences across groups in the observed test (i.e., subtest) scores. In other words, groups with the same latent (factor) means should achieve the same subtest scores. These conclusions would be supported by the maintenance of good model fit with the addition of intercept invariance constraints. If good model fit is not maintained, however, the differences in latent means would be misleading, and potential reasons for this lack of fit should be investigated (Keith & Reynolds, 2012).

Given a significant decline in model fit at the metric or intercept invariance level, partial invariance can be explored. At this level of analysis, focus was on bias related to specific subtests, supporting the assertion that establishing partial intercept invariance is akin to assessing content bias at the subtest level—as compared to content bias at the item level, which is commonly conducted during measurement development (e.g., Elliot, 2007). To assess for partial intercept invariance, modification indices were used to identify specific subtests that may account for the significant decrease in model fit. The measured intercepts for these subtests were systematically allowed to vary across groups, and model fit was re-evaluated with each adjustment. Adjustments to the model stopped when model fit did not degrade significantly, as this evidence suggested that partial metric/intercept invariance had been established. In this way, metric/intercept invariance could be recognized in the areas where it was present, and any specific problematic

subtests could be identified as such. It should be noted that it is possible to test for partial invariance at any level if complete invariance is not achieved (Keith & Reynolds, 2012); however, because intercept invariance is the strongest level of invariance testing commonly achieved, investigating partial invariance at the intercept invariance level is more common than at the preceding, weaker levels of testing.

Model fit. As mentioned previously, model fit was assessed at each level of invariance testing, and additional model fit comparisons were made at the metric and intercept invariance levels. Several fit indices are available for the purposes of evaluating model fit, with different fit indices incorporating varying aspects of fit. Consequently, the use of several criteria to establish model fit is recommended (Hu & Bentler,1999).

For assessing the fit of single models in the current study, the root mean square error of approximation (RMSEA), the comparative fit index (CFI), and the standardized root mean square residual (SRMR) were used (e.g., Boomsma, 2000; McDonald & Ho, 2002). RMSEA involves the lack of fit of the model in question compared to the population covariance matrix and is designed to measure approximate fit of a model per degree of freedom while taking sample size into account (Kline, 2005). An adjusted RMSEA based on the number of groups included in the analysis is recommended for use with multiple samples (Steiger, 1998) and was employed in the current study for group comparison analyses. Supported as an acceptable measure of overall model fit (Fan, Thompson, & Wang, 1999), RMSEA values below .05 indicate good fit (Browne & Cudeck, 1993; Hu & Bentler, 1999). Browne and Cudeck (1993) also suggested that values below .08 represent adequate fit, whereas values greater than .10 indicate poor fit.

The CFI calculates model fit based on comparisons between the model in question and the null, or baseline model, for which population covariances among observed variables are assumed to be zero (Kline, 2005). In general, CFI values approaching 1.00 indicate better fit, with values over .95 suggesting good fit, and values over .90 representing reasonable fit (Hu & Bentler, 1999). The SRMR represents the "average difference between the actual correlations among measured variables and those predicted by the model" (Keith, 2006, p. 270). Identified as one of the best measures of model fit, SRMR values below approximately .08 indicate good model fit (Hu & Bentler, 1998, 1999).

Model fit comparisons are necessary to determine whether the additional constraints or restrictions imposed at each level of invariance testing significantly degrade the fit of the model to the data. Determining the degree of invariance between two nested models is often assessed using the Likelihood Ratio Test—the difference in chi-square between two models ($\Delta\chi^2$; Cheung & Rensvold, 2002). However, evidence suggests that the χ^2 statistic is highly sensitive to sample size (e.g., Brannick, 1995; Kelloway, 1995) and that it may not be a practically useful measure for determining overall model fit or for comparing competing models when testing for invariance (e.g., Cheung & Rensvold, 2002). Several other fit indices have been proposed as practical alternative measures for overall fit, such as those mentioned above and used for the purposes of the current study, as well as alternative methods for comparing competing models. Based on simulation research, Cheung and Rensvold (2002) suggested the use of the CFI difference across models (across each level of invariance: ΔCFI). Strengths of both CFI and ΔCFI are that neither is affected by sample size, they are not significantly

correlated with each other, and neither is affected by model complexity (Cheung & Rensvold, 2002). This criteria has become common in CFA invariance studies (e.g., Weiss, Keith, Zhu, & Chen, 2013). The critical value for ΔCFI is -0.01, meaning that a decrease in CFI of greater than 0.01 across models suggests a lack of invariance across groups at that level and that bias in some form may be present.

Results

Initial Analyses

Initial models were analyzed with each individual sample prior to making any comparisons across groups in order to evaluate overall model fit with each individual sample. Subtest means and standard deviations for each sample are included in Table 3. The goodness-of-fit indices resulting from these analyses are included in Table 4 and support adequate to good model fit for all samples. With CFI values equal to or greater than .95 indicating good model fit (Hu & Bentler, 1999), results of these preliminary analyses (with a minimum CFI value of .97) suggest good model fit across all individual samples. Values for RMSEA below .05 indicate good fit, whereas values less than .08 indicate adequate fit (Browne & Cudeck, 1993; Hu & Bentler, 1999). Given these criteria, the RMSEA values for the Asian, Hispanic and both White samples suggest good fit, and for the Black sample (RMSEA=.053) suggests adequate to good fit. It should be noted that multi-sample adjustments in RMSEA were not made for these initial analyses, as there was only one sample analyzed at a time. Lastly, SRMR values below .08 suggest good model fit, and with SRMR values across all samples having values less than .062, further evidence was provided for good model fit. Based on these results, it was appropriate to move forward with the analyses without modifying the model in any way for any group.

Table 3

Group Subtest Means and Standard Deviations

	Asian		Black		Hispanic		White 1		White 2	
Subtest	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
Word Definition	53.00	9.55	46.39	9.00	47.06	8.73	51.30	9.60	51.49	9.92
Verbal Similarities	52.88	8.85	46.82	10.18	46.51	9.84	52.09	8.43	51.78	9.38
Matrices	56.68	10.58	46.60	10.47	48.30	8.97	51.29	9.75	50.93	9.80
Seq & Quant Reasoning	56.62	11.26	45.79	9.07	48.22	8.95	51.21	9.36	51.38	10.35
Pattern Construction	55.73	9.98	43.85	10.08	49.56	9.69	52.30	9.83	52.32	10.93
Recall of Designs	54.39	7.90	45.63	9.82	48.92	9.37	51.26	9.57	50.92	9.70
Recall of Obj Delayed	52.22	8.44	47.23	10.15	49.63	9.47	50.17	9.45	50.52	10.22
Recall of Obj Immediate	50.12	10.10	45.78	11.86	47.66	10.64	50.15	10.20	49.66	10.99
Digits Forward	55.32	13.04	49.65	10.42	46.70	10.16	50.48	11.19	49.98	10.96
Digits Backward	54.48	10.12	46.43	10.52	48.21	9.64	51.02	8.99	49.89	9.55
Recall of Seq Order	52.22	8.61	46.46	10.19	46.71	10.03	51.68	9.45	50.63	10.07
Speed of Info Processing	53.73	9.10	50.39	11.50	49.10	9.43	49.85	10.06	50.84	10.06
Rapid Naming	52.64	9.02	48.24	10.58	48.44	9.03	50.94	9.52	50.36	10.70

Table 4

Model Fit For Individual Samples.

Group	CFI	RMSEA	SRMR
Asian	.992	.027	.0614
Black	.971	.053	.0406
Hispanic	.980	.042	.0299
White 1	.976	.047	.0334
White 2	.991	.032	.0247

Data Analysis.

Asian. ***Initial comparisons: Asian-White 1.*** *Step 1: Configural invariance*. As discussed previously, to examine configural invariance—to assess whether the measure's factors have the same configuration across groups—the factors and the pattern of loadings (both free and fixed) were set to be the same for both groups being compared. Means of all latent variables were fixed to zero, while the intercepts for the measured variables were freely estimated for all groups. Measures of model fit, presented in Table 5, were examined before moving to the next level of invariance testing. Results of the initial comparison between the Asian group and White (1) group examining configural invariance suggest excellent model fit, with CFI and TLI values above the .95 cutoff (.979 and .967, respectively), an adjusted RMSEA value below .05 ($RMSEA_{adj}$=.044), and a SRMR value well below .08 (SRMR=.0334). This finding suggests that the factor model appears to be a good fit across both groups and that the factor configuration is similar for both groups.

Step 2: Metric invariance (weak factorial invariance). To assess for metric invariance, unstandardized factor loadings were constrained to be equal to corresponding factor loadings across both groups. As in the first level of invariance testing, measures of model fit were again considered, although at this level comparisons of model fit were also made. These measures of model fit, presented in Table 5, again suggested excellent fit between the model and the data: CFI and TLI values were both above the .95 cutoff for good fit (.971 and .958, respectively), the adjusted RMSEA was below .05 ($RMSEA_{adj}$=.049), and the SRMR (.0369) was well below the .08 cutoff. Although these values suggested a general decline in model fit as compared to those values associated with configural invariance, they are all still within the range of good fit. As mentioned previously, to determine whether there was a significant decline in model fit, the ΔCFI was considered, with a ΔCFI value greater than -0.01 indicating a significant decline in model fit between the previous level of invariance and the current level of invariance being considered (with additional equality constraints established). Additional constraints imposed to assess metric invariance between the Asian group and the White 1 sample resulted in a ΔCFI of -0.008, which is below the -0.01 cutoff and suggests that model fit did not significantly decline and provides support for adequate metric invariance between these two groups. In other words, it appears that the scale— or the relationship between the factors and the subtests—is the same for both groups. Having established metric invariance, intercept invariance could then be considered.

Step 3: Intercept invariance (strong factorial invariance). In order to determine whether there appears to be intercept invariance across these groups, factor means were

allowed to vary across groups and all corresponding subtest intercepts were constrained to be equal across groups. Overall model fit was again considered. With a CFI value above the .95 cutoff for good fit (CFI=.962), a TLI value above the .90 cutoff for adequate fit (TLI= .948), an adjusted RMSEA below the .08 cutoff for adequate fit ($RMSEA_{adj}$=.055), and SRMR below .08 (SRMR=.0387), overall model fit was still adequate to good. Additionally, the ΔCFI was below the -0.01 cutoff, suggesting a non-significant decline in model fit. These findings provide support for the presence of intercept invariance between the Asian and White 1 samples, which suggests that differences in latent (factor) means appear to account adequately for all mean differences on subtest scores.

Replication comparisons: Asian-White 2. The same steps as those described above and in the previous section were followed for all remaining analyses. In order to avoid redundancy, details regarding each step of analysis are only included if they deviate from those described previously.

Step 1: Configural invariance. Results of the comparison between the Asian group and White (2) group examining configural invariance suggest excellent model fit, with CFI and TLI values above the .95 cutoff (.991 and .986, respectively), an adjusted RMSEA value below .05 ($RMSEA_{adj}$=.031), and a SRMR value well below .08 (SRMR=.0247). These findings suggests that the factor model appears to be a good fit across both the Asian and White 2 groups and that the factor configuration is similar for both groups. Given the consistency of these results to those found between the Asian sample and the initial comparison group (White 1), there appears to be further support for

the presence of similar factor configuration for the Asian minor group as compared to the White major group.

Step 2: Metric invariance (weak factorial invariance). Measures of model fit at the metric invariance level across the Asian and White 2 groups, presented in Table 5, again suggest excellent fit between the model and the data: CFI and TLI values were both above the .95 cutoff for good fit (.987 and .982, respectively), the adjusted RMSEA was below .05 ($RMSEA_{adj}$=.037), and the SRMR was well below the .08 cutoff (SRMR=.0261). Regarding the comparisons between configural invariance and metric invariance, there was not a significant decline in model fit (ΔCFI=-.005, cutoff=-.01), which suggests that there is evidence for metric invariance across the Asian and White 2 groups. In other words, it appears that the test measures the same constructs for both groups. As with the replication of configural invariance between these groups, these results at the metric invariance level between the Asian and White 2 samples are also consistent with initial results between the Asian and White 1 samples, providing further evidence for the presence of metric invariance on the DAS-II for these groups.

Step 3: Intercept invariance (strong factorial invariance). Intercept invariance between the Asian and White 2 samples was also evaluated. With CFI and TLI values above the .95 cutoff for good fit (.962 and .977, respectively), an adjusted RMSEA below .05 ($RMSEA_{adj}$=.041), and SRMR below .08 (SRMR=.0266), overall model fit was excellent. Additionally, the ΔCFI was below the -.01 cutoff (ΔCFI=-.004), suggesting a non-significant decline in model fit. These findings support the presence of intercept invariance between the Asian and White 2 samples. Again, given their similarity with the

findings from the comparisons between the Asian and White 1 samples, the findings provide further support for the presence of intercept invariance across these groups.

Black. As with the replication process for the Asian sample, the same steps as those described above and in the previous section were followed for all remaining analyses. In order to avoid redundancy, details regarding each step of analysis will only be included if they differ in any way from those described previously. Measures of model fit and comparison values for all analyses are included in Table 5.

Initial comparisons: Black- White 1. *Step 1: Configural invariance.* Results of the initial comparison between the Black group and White (1) group examining configural invariance suggested excellent model fit, with CFI and TLI values above the .95 cutoff (.974 and .959, respectively), an adjusted RMSEA value below .05 ($RMSEA_{adj}$=.049, and a SRMR value well below .08 (SRMR=.0334). Thus, the factor model appears to be a good fit across both groups and the factor configuration appears to be similar for both groups.

Step 2: Metric invariance (weak factorial invariance). Overall measures of model fit at the metric invariance level also suggest good fit between the model and the data: CFI and TLI values were both above the .95 cutoff for good fit (.967 and .952, respectively), the adjusted RMSEA was below .08 ($RMSEA_{adj}$=.054), and the SRMR was well below the .08 cutoff (SRMR=.0419). Although these values suggest a general decline in model fit as compared to those values associated with configural invariance, they were all still well within the range of good fit. Additional constraints imposed to assess metric invariance between the Black group and the White 1 sample resulted in a

ΔCFI of -.009, which is below the -.01 cutoff and suggests that model fit did not significantly decline and provides support for adequate metric invariance between these two groups.

Step 3: Intercept invariance (strong factorial invariance). Regarding intercept invariance between the Black and White 1 groups, model fit was in the adequate to good range with a CFI value above the .95 cutoff for good fit (CFI=.956), a TLI value above the .90 cutoff for adequate fit (TLI= .939), an adjusted RMSEA below the .08 cutoff for adequate fit ($RMSEA_{adj}$=.061), and SRMR below .08 (SRMR=.0387). The ΔCFI at this level across the Black and White 1, however, was just over the -.01 cutoff with a ΔCFI value of -.011, suggesting a significant decline in model fit and evidence for potential bias at this level.

Consequently, further investigation into these differences was made, and partial intercept invariance was explored. Based on the modification indices provided, the equality constraint across groups for the intercept of the Digits Forward subtest appeared to be most problematic; thus, analyses were run again with intercepts for this single subtest allowed to vary freely across groups. Observed means on the Digits Forward subtest were 50.48 for the White 1 group and 49.65 for the Black group, with a difference in means of .83. Intercept estimates for Digits Forward were 54.223 for the Black group and 50.481 for the White 1 group. Measures of overall model fit with this adjustment for partial intercept invariance were slightly better than those achieved with full intercept invariance, with a CFI of .963 (above the .95 cutoff for good model fit), a TLI of .948 (above the .90 cutoff for adequate model fit and approaching the .95 cutoff for good fit),

adjusted RMSEA of .057 (below the .08 cutoff for good to adequate fit), and a SRMR of .0382 (well below the .08 cutoff for good fit). The ΔCFI from the metric invariance model to the partial intercept invariance model was -.007, which is below the -.01 critical value and suggests that model fit did not significantly decline with partial intercept invariance with the Digits Forward subtest allowed to vary. These findings support the presence of partial intercept invariance between the Black and White 1 groups, meaning that the Digits Forward subtest may show differences across groups that are not accounted for by differences in the latent Gsm (working memory) factor. The remaining subtests appear to have adequate intercept invariance across groups.

Replication comparisons: Black- White 2. *Step 1: Configural invariance*. Results of the comparison between the Black group and White (2) group examining configural invariance suggest excellent model fit, as CFI and TLI values were above the .95 cutoff (.982 and .973, respectively), the adjusted RMSEA was below .05 ($RMSEA_{adj}$=.044), and the SRMR was well below .08 (SRMR=.0247). These findings suggests that the factor model appears to be a good fit across the Black and White 2 groups and that the factor configuration is similar for both groups. Given the consistency of these results to those found between the Black sample and the initial comparison group (White 1), there appears to be further support for the presence of similar factor configuration for the Black minor group as compared to the White major group on the DAS-II.

Step 2: Metric invariance (weak factorial invariance). Measures of model fit at the metric invariance level across the Black and White 2 groups also suggest excellent fit

between the model and the data. CFI and TLI values were both above the .95 cutoff for good fit (.982 and .973, respectively), the adjusted RMSEA was below .05 ($RMSEA_{adj}$=.047), and the SRMR value was well below the .08 cutoff (SRMR=.0247). Regarding the comparisons between configural invariance and metric invariance, there was not a significant decline in model fit (ΔCFI=-.004, cutoff=-.01), which suggests that there is evidence for metric invariance across the Black and White 2 groups. As with the replication of configural invariance between these groups, these results at the metric invariance level between the Black and White 2 samples are also consistent with initial results between the Black and White 1 samples, providing further evidence for the presence of metric invariance on the DAS-II across these groups.

Step 3: Intercept invariance (strong factorial invariance). Regarding intercept invariance between the Black and White 2 samples, overall model fit was good. With CFI and TLI values above the .95 cutoff for good fit (.972 and .961, respectively), an adjusted RMSEA below .08 and approaching the .05 cutoff ($RMSEA_{adj}$=.051), and a SRMR below .08 (SRMR=.0306), model fit for these groups at this level is better than model fit for the Black and White 1 groups at this level. Additionally, the ΔCFI was below the -.01 cutoff (ΔCFI=-.006), suggesting a non-significant decline in model fit. These findings support the presence of intercept invariance between the Black and White 2 samples.

Hispanic. ***Initial comparisons: Hispanic-White 1.*** *Step 1: Configural invariance*. The initial comparison between the Hispanic group and White (1) group examining configural invariance suggest excellent model fit, with CFI and TLI values

above the .95 cutoff (.978 and .966, respectively), an adjusted RMSEA value below .05 ($RMSEA_{adj}$=.045), and a SRMR value well below .08 (SRMR=.0334). Thus, the factor model appears to be a good fit across both groups and the factor configuration appears to be similar for both groups.

Step 2: Metric invariance (weak factorial invariance). Measures of overall model fit suggest excellent fit between the model and the data for the Hispanic and White 1 groups at the metric invariance level as well: CFI and TLI values were both above the .95 cutoff for good fit (.976 and .964, respectively), the adjusted RMSEA was below .05 ($RMSEA_{adj}$=.045), and the SRMR was well below the .08 cutoff (SRMR=.0364). Additional constraints imposed to assess metric invariance between the Hispanic group and the White 1 sample resulted in a ΔCFI of -.002, which is below the -.01 cutoff and suggests that model fit did not significantly decline and provides support for adequate metric invariance between these two groups.

Step 3: Intercept invariance (strong factorial invariance). At the level of intercept invariance, overall model fit remained good, with CFI and TLI values above the .95 cutoff for good fit (CFI=.972, TLI=.961), an adjusted RMSEA below .05 ($RMSEA_{adj}$=.048), and SRMR below .08 (SRMR=.0366). Additionally, the ΔCFI was below the -.01 cutoff (ΔCFI =-.004), suggesting a non-significant decline in model fit. These findings provide support for the presence of intercept invariance between the Hispanic and White 1 samples.

Replication comparisons: Hispanic-White 2. *Step 1: Configural invariance*. Regarding the comparison between the Hispanic group and White (2) group at the

configural invariance level, overall measures of model fit suggest excellent model fit, with CFI and TLI values above the .95 cutoff (.986 and .979, respectively), an adjusted RMSEA value well below .05 ($RMSEA_{adj}$=.038), and a SRMR value well below .08 (SRMR=.0247). These findings suggests that the factor model appears to be a good fit across both the Hispanic and White 2 groups and that the factor configuration is similar for both groups. Given the consistency of these results to those found between the Hispanic sample and the initial comparison group (White 1), there appears to be further support for the presence of similar factor configuration for the Hispanic minor group as compared to the White major group.

Step 2: Metric invariance (weak factorial invariance). Overall measures of model fit at the metric invariance level across the Hispanic and White 2 groups, again suggest excellent fit between the model and the data: CFI and TLI values were both above the .95 cutoff for good fit (.986 and .980, respectively), the adjusted RMSEA was below .05 ($RMSEA_{adj}$=.037), and the SRMR was well below the .08 cutoff (SRMR=.0254). Regarding the comparisons between configural invariance and metric invariance, there was not a significant decline in model fit (ΔCFI=-.000, cutoff=-.01), which suggests that there is evidence for metric invariance across the Hispanic and White 2 groups. As with the replication of configural invariance between these groups, these results at the metric invariance level between the Hispanic and White 2 samples are also consistent with initial results between the Hispanic and White 1 samples, providing further evidence for the presence of metric invariance on the DAS-II for these groups.

Step 3: Intercept invariance (strong factorial invariance). With CFI and TLI values above the .95 cutoff for good fit (.984 and .978, respectively), an adjusted RMSEA value below .05 ($RMSEA_{adj}$=.038), and SRMR value below .08 (SRMR=.0258), overall model fit for the intercept invariance model for the Hispanic and White 2 samples was excellent. Additionally, the ΔCFI was below the -.01 cutoff (ΔCFI=-.002), which is indicative of a non-significant decline in model fit. These findings support the presence of intercept invariance between the Hispanic and White 2 samples. Again, given their similarity with the findings from the comparisons between the Hispanic and White 1 samples, the findings provide further support for the presence of intercept invariance across these groups on the DAS-II.

Table 5

Summary of Fit Indices for MG-CFA Analyses

Groups	**Invariance Level**	**CFI**	**ΔCFI**	**TLI**	**RMSEA**	**SRMR**
Asian- White1	Configural	.979		.967	.044	.0334
	Metric	.971	.008	.958	.049	.0369
	Intercept	.962	.009	.948	.055	.0387
Asian- White2	Configural	.991		.986	.031	.0247
	Metric	.987	.005	.982	.037	.0261
	Intercept	.983	.004	.977	.041	.0266
Black- White 1	Configural	.974		.959	.049	.0334
	Metric	.967	.009	.952	.054	.0364
	Intercept	.956	.011*	.939	.061	.0419
	Partial-Intercept	.963	.007	.948	.057	.0382
Black- White 2	Configural	.982		.973	.044	.0247
	Metric	.978	.004	.968	.047	.0270
	Intercept	.972	.006	.961	.051	.0306
Hispanic-White 1	Configural	.978		.966	.045	.0334
	Metric	.976	.002	.964	.045	.0364
	Intercept	.972	.004	.961	.048	.0366
Hispanic-White 2	Configural	.986		.979	.038	.0247
	Metric	.986	.000	.980	.037	.0254
	Intercept	.984	.002	.978	.038	.0258

Note. *Denotes significant change in model fit

Discussion

The DAS-II (Elliott, 2007) is a relatively new, individually administered test of cognitive abilities with growing popularity. Consequently, it is essential that there be empirical evidence supporting the use of the DAS-II as an appropriate measure of cognitive abilities for children of varying backgrounds. In an effort to ensure that the measure met adequate reliability and validity criteria, the test publishers conducted extensive research with a representative sample during test development; however, the issue of test bias across racial/ethnic groups was not specifically explored. The original DAS has been evaluated for test bias across racial/ethnic groups (Keith, Quirk, Schartzer, and Elliott, 1999), but there were significant changes made from the original version to the second edition, and thus there is no guarantee that the evidence from the original would necessarily apply to the newer edition.

The purpose of the current study was to investigate whether the DAS-II demonstrates systematic construct bias toward children of any of the following racial/ethnic groups: Asian, Black, Hispanic, and White. More specifically, the current study aimed to address three primary research questions: (a) Is construct bias present in the DAS-II toward any of these groups? (b) If so, where does this bias exist? and (c) Can these findings be replicated with a second comparison group? In order to fulfill this purpose, multi-group confirmatory factor analysis was used to test for measurement invariance (or construct bias) across groups using data from the DAS-II standardization sample. This methodology incrementally tested whether criteria for increasingly strict

levels of invariance were met across groups (e.g., Keith & Reynolds, 2012; Meredith, 1993).

Overall, the results of these analyses provide strong support for construct validity and a lack of construct bias for Asian and Hispanic children and youth, as compared to White children and youth. In other words, the underlying attributes and constructs measured by the DAS-II appear to be similar, or at least not statistically discernible, across these groups. Results from the current study regarding Black children and youth, however, are somewhat less clear due to somewhat inconsistent findings between the initial comparison group (White 1) and the replication sample (White 2).

When the Black group was compared to the White 1 group at the configural and metric invariance levels, results supported the presence of both across groups, suggesting that the configuration of subtests and factors is consistent across groups, and the scale for the measure is consistent across groups. It appeared, however, that full intercept invariance might not be tenable due to a decrease in model fit that was slightly greater than the cutoff value ($\Delta CFI > .01$). Thus, partial intercept invariance was explored and established across these two groups by allowing the intercept of the Digits Forward subtest to vary across groups. The significant improvement in model fit associated with this change suggests that there may be evidence of bias associated with this subtest. In other words, it is possible that the Digits Forward subtest is differentially more difficult for one group over another, even when controlling for latent short-term memory abilities.

If confirmed, these findings could have problematic implications for two reasons. First, the subtest shows higher intercept scores for one group (Black) when those same

differences do not exist on the underlying factor or ability. This inconsistency may be indicative of a problem with the subtest and may suggest a need for caution when interpreting results on this subtest, as the subtest may not accurately reflect abilities for these children (i.e., it could potentially over-estimate scores for Black children). Second, problems with intercept invariance may also result in non-valid group differences on composite scores that are created using the subtest in question. The Digits Forward subtest is a diagnostic subtest and contributes only to the diagnostic cluster for working memory, which suggests that this cluster score may also require interpretation with caution when used with Black children and youth. It should be noted that the Digits Forward subtest is not used in the calculation of the (overall) general cognitive ability score for the DAS-II, and thus this composite is not affected by these findings.

Despite these important considerations, these findings should be interpreted with caution for several reasons. To begin, there does not appear to be an apparent cultural or environmental explanation available for these findings. Moreover, these results were not consistent with findings from comparisons with the replication sample. When compared to the White 2 group, the Black group was statistically indistinguishable at any level of invariance, thus providing good support for measurement invariance or a lack of construct bias across these groups. In fact, given that the comparisons between the Black and the initial White (1) groups maintained adequate to good model fit across all models and experienced a change in model fit at the intercept invariance level right at the cusp of significance (.01 cutoff vs .011 value), and considering the excellent model fit and support for measurement invariance across the Black and White 2 groups, it appears that

there is still adequate evidence for a lack of construct bias across these groups. In short, although the findings are not as robust as with the other groups, it appears to be a reasonable conclusion that the DAS-II is not likely biased in favor of or against Black children and youth. Thus, to summarize, the results of the current study suggest that the subtests from the DAS-II included in the current study do not appear to show evidence of construct bias across any of the groups included in the current study.

Limitations

Despite the general conclusion that there does not appear to be evidence of construct bias on the DAS-II across the groups in the current study, there are several considerations that need to be made in order to best understand the meaning of these findings in a greater context. For one, limitations in the current study related to the measure itself as well as the methodology are important to acknowledge and address through future research. First, pertaining to the measure itself, the current study included only those subtests administered to all children in the school-age sample. Further investigation into test bias with the DAS-II should consider the remaining school age subtests as well as subtests comprising the younger age battery in order to ensure that the evidence for a lack of construct bias also applies to these subtests and age groups. Also, given the hierarchical organization of the subsets of validity (as discussed in Appendix) and that there has been investigation into content bias at the item level (Elliott, 2007) and now construct bias (current study), predictive bias for the DAS-II will be important to examine as well. There is some preliminary evidence for predictive validity of the DAS-II (Elliott, 2007); however, whether there is consistent predictive validity across

racial/ethnic groups (predictive bias) has not yet been explored. Additionally, although the ability to assess the reliability of the findings by replicating the initial analyses with a second comparison group can be considered a strength of the current study, the results of this validation process would be strengthened further had new minor group samples been incorporated in addition to the major group comparison sample. Conducting future research that addresses these issues would further bolster evidence for the absence of bias in the DAS-II across groups.

It is also important to acknowledge the limitations regarding the generalizability of these results. For one, the generalizability of these results may be limited due to factors related to the data available and the demographics of the participants. The samples used for the current study were obtained from the overall standardization sample for the measure, which was stratified based on the U.S. Census Bureau's Current Population Survey from 2005. Needless to say, the demographics of the United States have changed significantly since that time, which suggests that the sample may no longer be representative in the way it was previously. Moreover, the lack of details in the manual regarding the country of heritage is problematic, as labels such as "Asian" or "Hispanic" (those used in the DAS-II technical manual and used to determine group membership) typically include individuals with backgrounds from a number of countries with diverse histories, cultures, and languages, etc. which would imply limited generalizability of results that do not include representative subsamples from each country. Similarly, there is tremendous cultural variability across states in the US, and given the lack of details regarding representation by state in the standardization sample, one cannot be entirely

confident that the findings of the current study are applicable across all states in the US. A further problem exists regarding the lack of details pertaining to methods used to determine language dominance in the DAS-II technical manual (Elliott, 2007), as well as the lack of details regarding proficiency levels in English for participants. One cannot be sure that all children who participated in the standardization of the measure are in fact English dominant or proficient in English to a certain identifiable degree.

Additionally, only four broad racial/ethnic groups were included in the current study. Although it is somewhat unusual in studies using standardization data to have an Asian sample large enough to include in test bias analyses (what might be considered a strength of the current study as compared to the test bias literature overall- see Valencia & Suzuki, 2001, Chapter 5), the current study did not include additional racial/ethnic groups such as Native American or Pacific Islander. It follows then that the evidence of non-bias in the current study may not necessarily apply to other racial or ethnic groups that were not included in the study. These limitations not only contribute to potentially limited generalizability of the results of the current study, but also represent ways in which standards for research in the domain of test bias should be raised. As an aside, they additionally pertain to test developers as important considerations to be made during the process of identifying participants to include as part of the standardization samples used to norm assessment measures.

This point leads to recognition of further limitations for the current study and test bias research as a whole. For example, considerations such as how racial/ethnic groups are defined or whether racial/ethnic categories should be used as explanatory constructs

at all (e.g., Helms, Jernigan, & Mascher, 2005) are bigger, more conceptual questions that are directly related to this type of research. Additionally, although the current study provides evidence that the DAS-II measures the same underlying constructs (that has a statistically similar configuration, scale, and equivalent starting levels given the same latent abilities) across groups, the current study does not address whether this underlying construct does, in fact, represent a "true" or an unbiased definition of intelligence or cognitive abilities.

Implications

Outcomes of this research serve to support the general appropriateness of the DAS-II for clinical use with several racial/ethnic groups. Psychologists and those administering and interpreting scores from this measure can feel reasonably confident that the construct measured by the DAS-II is likely equivalent for the ethnic/racial groups considered in the current study. They can also feel reasonably confident that results from the DAS-II, when obtained through a standard administration and considered in conjunction with relevant data from other sources, can provide useful information regarding the cognitive performance of children and youth. As with all cognitive measures, however, it is important to recognize that a single number does not capture the full range of a person's abilities, but rather represents an approximation of that person's performance on a specific task in a certain context.

A continued word of caution regarding the interpretation of cognitive test results pertains to the significance and implications of group mean differences on these measures, which is a noteworthy source of controversy that continues to fuel the test bias

debate. This consideration is particularly relevant under circumstances similar to those in the current investigation—where evidence suggests a lack of construct bias in the measure across racial/ethnic groups—as some may misinterpret the significance or implications of mean score differences on a "non-biased" measure. The meaning of group mean differences on this measure is limited for one by the factors identified in the previous section. Second, for example, some elect to interpret group mean differences on "unbiased" cognitive measures as evidence for genetically inferior intellect of those groups scoring lower on these measure (see e.g., Appendix); however, this perspective is largely considered to be rooted in pseudoscience and fails to consider several alternative explanations for group differences, such as opportunity to learn, language, socioeconomic status, and segregated schools (e.g., Valencia, Rankin, & Livingston, 1995; Valencia, 2010; Valencia, 2013). Although a more detailed discussion of these issues is beyond the scope of the current study, these points serve to support the need to a) consider results from the current investigation in a greater context, b) conduct further research in the area of test bias, and c) reconsider the methods and standards used to conduct test bias research.

Appendix: Literature Review

The following literature review begins with an overview of intelligence testing, including a brief history of its development and application, a discussion of the uses of intelligence tests, and a brief discussion of some of the issues associated with intelligence testing. As will become evident in this overview, bias in intelligence testing is both a controversial topic and an imperative issue to be addressed. A discussion of bias in intelligence testing follows, with an exploration into what *is not* meant by bias, what *is* meant by bias, and how bias can be and has been measured. This section concludes with a brief review of recent research investigating bias in the testing literature.

Overview

Historical development. Before one is able to fully understand the current issues surrounding intelligence testing, particularly the issue of test bias, one must first consider the historical context and development within the field of intelligence testing. Although questions regarding mental ability and how to measure it have been asked all over the world for hundreds of years (e.g., Cohen & Swerdlik, 2002), our contemporary model of intelligence assessment is generally considered to have its roots in the emergence of the intelligence testing movement in Europe that began in the late 19th century.

Sir Francis Galton. Sir Francis Galton is generally credited with pioneering the psychological testing movement (e.g., Anastasi, 1996; Cohen & Swerdlik, 2002; Sattler, 2001). Following Charles Darwin's discussion of individual differences in his publication of *On the Origin of Species by Means of Natural Selection* (1859), Sir Francis

Galton developed an interest in heredity and subsequently published his book *Hereditary Genius: An Inquiry into its Laws and Consequences* (1869). Within this text, Galton presented his "pioneering" views of intelligence, arguing that, "a man's natural abilities are derived by inheritance, under exactly the same limitations as are the form and physical features of the organic world" (p. 1), and suggesting that individual and group differences, in several domains—particularly intelligence—are primarily attributable to genetics.

As one might infer, Galton's work can be connected to the modern-day nature vs. nurture debate, which influenced the development of intelligence theory in its own right, but he is often credited with initiating the psychological testing movement for additional reasons (Anastasi, 1996; Cohen & Swerdlik, 2002; Sattler, 2001). In his book, *Inquiries into Human Faculty and its Development* (1883), Galton described several techniques and methods he used— predominantly physical and sensory measurements— to "take note of the varied hereditary faculties" (p. 1) which he believed were closely connected to the "innate moral and intellectual faculties" of individuals and groups. In this compilation of articles, he also offered some consideration for the problems involved in measuring mental characteristics. Although it was Pearson who developed the product-moment correlation technique, the roots of this concept can be linked to Galton's work, and Galton is ultimately recognized as the person who developed the central statistical concepts of regression to the mean and the coefficient of correlation (Sattler, 2001).

Alfred Binet and Théodore Simon. Although Galton is credited with initiating the psychological testing movement, the development of the first cognitively based psychological measure is attributed to French psychologists Alfred Binet and Théodore Simon (1905). Binet and Simon (1895; cited in Anastasi, 1996) criticized existing measures of the time as being too highly specialized, too simple, and too focused on sensory perception, which contributed to the impetus for their development of the Binet-Simon Scale (Binet & Simon, 1905). As a cognitively-based measure, the Binet-Simon Scale included tests for memory, comprehension, attention, judgment, and reasoning, which to a large degree have become the basis for intelligence tests used today. In contrast to Galton's hereditary view of intelligence, Binet argued that intelligence is influenced by a number of interrelated processes and is subject to change over time (Siegler, 1992).

Henry H. Goddard and Lewis M. Terman. Henry H. Goddard introduced the Binet-Simon Scales to the United States in 1908 and is credited with being the first to translate the scales into English, make minor revisions, and standardize it with American children (Goddard, 1910). Despite the contradiction to Binet's view of intelligence, Goddard argued that the scale measured intelligence as a unitary function primarily determined by heredity (Blum, 1978; Gould, 1981; Tuddenham, 1962). Goddard's work provided the foundation for Lewis M. Terman's revisions to the scale. Terman made extensive modifications to the scale and published it in 1916 as the Stanford Revision and Extension of the Binet-Simon Scale (Stanford-Binet; Terman, 1916). Terman (1916) described several significant revisions made to the Binet-Simon scales, including 1)

lengthening the measure by including more subtests, 2) extending the upper limit of the scale to include the "superior adult" level, 3) incorporating Stern's *mental quotient* (calculated by dividing mental age by chronological age; see Stern, 1914), 4) establishing a standard deviation of 15 to 16 points at each age level, 5) scaling the test to approximate a normal, bell-shaped curve, and 6) standardizing the scale with 1000 children in California (with emphasis placed on using subjects who were "nearly as possible representative of the several ages… in a community of average social status" (p. 52)). The introduction of the Binet-Simon Scale, as well as the revisions and methods employed by Terman to create the Stanford-Binet launched both the intelligence testing movement and the clinical testing movement and stimulated the development of these movements around the world. Tuddenham (1962) summarizes this influence:

> The success of the Stanford-Binet was a triumph of pragmatism, but its importance must not be underestimated, for it demonstrated the feasibility of mental measurement and led to the development of other tests for many special purposes. Equally important, it led to a public acceptance of testing which had important consequences for education and industry, for the military, and for society generally. (p. 494)

Although the tone of this excerpt is generally positive, Tuddenham provides a bit of foreshadowing through his mention of the "important consequences" of the public acceptance of testing, a point which will be discussed in more depth in a later section of this review.

David Wechsler. David Wechsler represents the next significant milestone in intelligence test development. His measure, the Wechsler-Bellevue Intelligence Scale (Wechsler, 1939a, 1939b), did not consist of any original tests; rather, it was a synthesis of previously existing tests and materials that gleaned virtually immediate popularity, as evidenced by positive reviews published upon its release (see, e.g., Boake, 2002; Buros, 1941, 1949; Lubin, Wallis, & Paine, 1971; Tulsky, 2003). Following the production of the Wechsler-Bellevue scale, Wechsler developed more specialized instruments, including the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949), the Wechsler Adult Intelligence Scale (WAIS; Wechsler, 1955), and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI; Wechsler, 1967). In the 1950's and 1960's, Wechsler's intelligence scales surpassed the Stanford-Binet as the most widely-used measure of intelligence (Lubin, Wallis, & Paine, 1971), and more recent surveys indicate that the widespread popularity of Wechsler's scale remains intact (eg. Archer, Maruish, Imhof, & Piotrowski, 1991; Camara, Nathan, & Puente, 2000; Lubin, Larsen, & Matarazzo, 1984; Wilson & Reschly, 1996).

Wasserman & Tulsky (2005) attribute the popularity of the Wechsler intelligence scales to 1) a lack of available intellectual assessment measures for adults, 2) the "rare" integration of verbal and performance tests into a single battery, 3) early validity research conducted through the "co-norming" with other tests that were commonly used in practice, 4) an exceptional norming sample for the time, and 5) special emphasis on "psychometric rigor." In sum, "the practice of intellectual assessment in the second half

of the 20th century may arguably been most strongly influenced by the work of David Wechsler" (1986-1981; Wasserman & Tulsky, 2005, p. 12).

Intelligence theory. Parallel to the intelligence testing movement came developments in intelligence theory and related statistical methods. One of the fundamental topics in question throughout the testing movement is how to define or conceptualize intelligence as a construct. This debate has already been introduced to a limited extent in the previous section's discussion of Galton, Binet, and Goddard; however, the debate continued long after, and after over 100 years of dialogue, there is still no consensus regarding a definition of intelligence (e.g., Sattler, 2001; Wasserman & Tulsky, 2005). A plethora of definitions of intelligence have been proposed over time, and with each new conceptualization, intelligence theory has evolved, and the question of how to measure intelligence has continued to be asked. These issues have also risen in relation to issues of test bias, as the underlying theories of intelligence have often influenced the nature of tests as well as their application. Over the course of time, statistical methods similar to those used to develop models of intelligence have also been used to assess bias within measures of intelligence. The next several sections include a brief review of theories of intelligence contributing to the CHC theory, which underlies the model of intelligence used for the current study (for a more detailed review of these and other contributing theories, see, e.g., McGrew, 2005).

Charles E. Spearman: g factor. Charles E. Spearman is credited with publishing one of the first attempts at formulating a theory of intelligence that was based on empirical foundation (Spearman, 1904). Spearman's research provided empirical support for one overarching factor of intelligence (g) and sparked controversy in the field for much of the 20th century. His original two-factor theory (1904) consisted of this overarching g factor, which was mathematically derived and accounted for the shared variance across intelligence tests, and specific s factors, which accounted for specific test-related variance. Despite intense criticisms, Spearman's g has been retained in most contemporary models of intelligence and has been described as "one of the most central phenomena in all of behavioral science, with broad explanatory powers" (Jensen, 1998, p. xii). In addition to the contribution of his g factor, Spearman is also credited with pioneering the statistical method that ultimately became factor analysis, variations of which are still employed in intelligence research today.

Louis L. Thurstone: Multiple-factor models of intelligence. More complex models were proposed as statistical methods become more sophisticated, with Louis L. Thurstone at the forefront of true factor analysis methods. Using these more advanced methods, Thurstone began in the mid-1930s to provide evidence counter to Spearman's general factor and to set the foundation for the multiple-factor models that characterize contemporary intelligence models (eg. Thurstone,1938). However, with time and the development of higher-order factor analytic techniques, Thurstone eventually conceded to the possibility of a general *g* factor (Thurstone, 1947).

Raymond B. Cattell and John L. Horn: Gf-Gc. Raymond B. Cattell used factor analytic techniques to develop a two-factor model of intelligence (*Gf-Gc* theory), which he introduced in 1941 (Cattell, 1941). Cattell argued that Spearman's *g* was insufficient in explaining intelligence and that intelligence could be better explained by a two-factor model, comprised of a "fluid ability" or "fluid intelligence" factor (*Gf*) and a "crystallized ability" or "crystalized intelligence" factor (*Gc*). John L. Horn, a student of Cattell's, joined Cattell in the 1960s in research that would expand the number of ability factors from 2 to 5 (Horn & Cattell, 1966). Horn later expanded the model further, proposing nine ability factors (e.g. Horn & Noll, 1997). Throughout the course of their research, Cattell and Horn argued against a general *g* factor, as they maintained their early argument that a single general factor did not fit the data.

Vernon, Guilford, and Gustafsson: Intermediate models. Three intermediate models were proposed between Cattell and Horn's *Gf-Gc theory* and John B. Carroll's Three-stratum model. Although these three models did not have the same degree of impact and are not nearly as well-known as the two models preceding and following them, it is important to acknowledge these intermediary steps. P. E. Vernon (1961) proposed a hierarchical factorial model that included a super-ordinate *g* factor and two lower-order factors, labeled "verbal-educational ability" (*v:ed*) and "mechanical-spatial ability" (*k:m*). These two lower-order factors were sub-divided further into two and three abilities, respectively. J.P. Guilford (1967), in his structure-of-intellect theory, rejected the verbal-nonverbal dichotomy, claiming that different abilities are involved when dealing with different kinds of information. He proposed four categories in the "content

of intellect": *figural, symbolic, semantic,* and *behavioral.* The third and final intermediate model of note was proposed by G. Gustafsson (1984), in which he formulated an integrated hierarchical model of intelligence. He proposed a general ability (*g*) at the highest level, with two broad factors at the next level: *crystallized intelligence* and *general visualization*, for verbal and figural information, respectively. This integrated hierarchical model is a significant precursor to later research, as Gustafsson and Vernon both anticipated a structure that would be integral for future models.

John B. Carroll: Three-stratum model. Following the contributions provided by Vernon, Guilford, and Gustafsson discussed in the previous section, John B. Carroll published his seminal work in the field, *Human Cognitive Abilities: A Survey of Factor Analytic Studies* (1993) in which he described and reported findings of a comprehensive meta-analysis of intelligence testing data for roughly 50 years. Through the implementation of thorough and systematic factor analyses, Carroll ultimately proposed a hierarchical, three-stratum (three level) model of intelligence. Within this model, Carroll included a third-order factor of general intelligence (*g*), eight (or more) second-order factors representing broad intellectual abilities, and 65 first-order factors, representing narrow intellectual abilities. Carroll's research provided evidence that the third-order *g* factor is most related to factors of variables involving reasoning ability, which suggests that the *g* factor of Carroll's model may be similar to (or even synonymous with) the *Gf* of Cattell and Horn's model.

Cattell-Horn-Carroll: CHC theory. CHC theory represents the integration of Cattell and Horn's Gf-Gc model (Horn & Noll, 1997) and Carroll's three-stratum model of intelligence (Caroll, 1993). This theory is considered a psychometric theory, as it is based on statistical methods that assume that "the structure of intelligence can be discovered by analyzing the interrelationship of scores on mental ability tests" through the creation and evaluation of models via factor analysis (Davidson & Downing, 2000, p. 37). McGrew (2005) conducted a detailed review of the CHC literature, including studies intentionally investigating various aspects of the CHC model and studies in which the CHC model was applied via post hoc analysis, and he concludes that the literature collectively provides support for the "broad strokes of contemporary CHC theory" (McGrew, 2005, p. 149). Moreover, McGrew indicates that, when compared to other popular theories of intelligence and cognitive abilities, CHC theory is the most comprehensive theory and has the most empirical support (2005). In a more recent review, Keith and Reynolds (2010) suggest that, although not without its limitations, CHC theory "offers the best current description of the structure of human intelligence" (p. 8). As such, CHC theory has been applied extensively in the development of various test batteries, as well as incorporated into the interpretation of existing batteries.

Applications and uses. As previously stated, the introduction of intelligence tests led to the widespread development and implementation of psychological testing. Not long after the Binet-Simon Scale was introduced, psychological tests were used in a wide range of settings, such as juvenile courts, reformatories, prisons, children's homes, and schools (Pintner, 1931). The Binet-Simon Scale helped "school systems identify

students who were likely to learn less rapidly than most of their peers and who, therefore, were candidates for special education" (Siegler, 1992, p. 185), which was its original intended purpose in France in the early 20th century. Use of cognitive tests was also common outside of schools, with one early example being the US military's administration of cognitive tests to all army recruits during World War I (Yerkes, 1921; Yoakum & Yerkes, 1920). Additional early applications for intelligence tests included evaluation of immigrants at Ellis Island to determine whether they were "mentally deficient" and thus not eligible for entry into the United States (Goddard, 1917). Early applications also included assessment of individuals in order to sort and categorize them for varying levels of job eligibility (i.e., unskilled labor based on measured IQ), testing of people believed to be "feeble-minded" in order to institutionalize them (as feeble-mindedness, or a lack of intelligence, was associated with a lack of morality and thus contributed to "crime, pauperism, and industrial inefficiency"), and evaluating those suspected of being "morons" with the goal of preventing them from having children and passing on their "gene for low intelligence" (Gould, 1996).

Although some of these uses would now be considered unacceptable by today's standards, many contemporary uses for and applications of cognitive testing are similar to those of earlier decades. As previously stated, Binet's measure was originally designed for the identification of students who would need special education in schools, and cognitive tests continue to serve this function. In addition to identifying children who have learning disabilities and might need special assistance in school, cognitive measures are also employed in the identification of gifted children for the purposes of tailoring

their education as well (e.g., McIntosh & Dixon, 2005). Classification also remains a broad objective of cognitive testing; however, the purposes and implications of this classification have changed dramatically since the early 20th century. Dowdy, Mays, Kamphaus, and Reynolds (2009) identified several functions of classification that apply to contemporary cognitive testing, including determining eligibility for special education services, identifying an individual's need for services, determining parameters (i.e., intensity and duration) of treatment, enhanced communication among professionals, ease of description, the necessity of valid taxonomies for grouping individuals for research, and the ability to differentiate abilities. Additional objectives of cognitive assessment include treatment planning, political advocacy, program evaluation, and research (Keough, 1994). In general, the administration of cognitive tests is common for a variety of purposes, across a wide array of settings, such as schools, clinics, hospitals, industry, and the military. Intelligence testing also influences public policy, business, and scientific psychology (e.g., Sattler, 2001), with, as quoted earlier, "consequences for education, industry, the military, and society generally" (Tuddenham, 1962, p 494). Sattler (2001) summarizes this abundance of testing uses and applications: "The testing movement, although subject to criticism, continues to thrive in the United States and in many other parts of the world." (p. 134).

Issues surrounding intelligence testing. Up to this point, this review has primarily been a summary of selected topics related to cognitive testing, including the historical development of the intelligence testing movement, highlights of the development of intelligence theory, and a description of the many uses and applications

of intelligence tests. Although presented in a relatively uncritical manner thus far, each of these areas has significant problems and criticisms relating to the issue of test bias, which is both a core question of the proposed study, as well as a fundamental concern in the domain of intelligence testing. Many of these issues will be addressed in more depth later; however, this portion of the review will provide a brief transition in order to bridge the previous sections that broadly address intelligence testing with the following sections that address the more specific issue of test bias.

The above survey of historical development in the intelligence testing movement provides a summary of the ideological underpinnings of the testing movement, identifies key figures in this process, and highlights the processes involved in launching the testing movement. Several things to consider, which will be discussed in more detail in the next section, are the potential ways in which each of these— the ideologies, the individuals, and the processes— could influence test development, application, and interpretation.

As established in the summary of intelligence theory, one of the biggest issues facing the field, from its inception to the present, is that there is a continued lack of consensus in the field regarding the definition of intelligence. There is so little consensus, in fact, that Jensen (1998) suggested that psychologists, "drop the ill-fated word from our scientific vocabulary, or use it only in quotes, to remind ourselves that it is not only scientifically unsatisfactory, but wholly unnecessary" (p. 49). Although this recommendation is a bit extreme, it illustrates the point that the term "intelligence" used as broadly and indiscriminately as it is, actually has limited value in itself. A related issue pertains specifically to the psychometric definitions of intelligence, as they underlie

the apparently circular manner in which theory influences measurement, which in turn influences theory. As Boring noted in 1923, intelligence is defined as whatever intelligence tests measure. This observation only becomes more relevant as models of intelligence are increasingly based on the data coming from intelligence tests. Implications for test bias are significant, especially with the reification of intelligence that occurs during the early 20th century (more on this will be presented in the next section).

Beyond the primary issue of defining intelligence is the secondary issue of the application, use, and interpretation of intelligence tests. As illustrated previously, these applications were (and still are) extensive, spanning a wide range of functions and settings. Questions of bias were raised shortly after intelligence tests began to be used. In fact, Binet is credited with conducting the first investigation into "cultural" bias around 1910 (Binet & Simon, 1916/1973) and was followed shortly after by Stern (1914). Although one might find this surprising given the trajectory that his test ultimately followed, Binet offered warnings regarding the use of his measure, identifying its limitations and indicating that longer-term influences, such as family and school background, health, and past effort in school could potentially affect test performance. In light of these limitations, Binet only recommended comparisons of test results among children from comparable backgrounds (Binet & Simon, 1916/73). Despite Binet's warnings, tests were administered to virtually anyone and everyone, often without consideration of the appropriateness of the measure, and frequently with important outcomes and decisions at stake.

Bias in Intelligence Testing

Questions of bias, particularly cultural bias, have been the source of great controversy since they were first raised, and they remain significant issues in contemporary professional literature and the popular press. The issue of bias was raised for several reasons, but one of the most fundamental reasons is likely connected to the differences in intellectual performance across groups, particularly across racial or ethnic groups. This discrepancy was established very early in the testing movement when Strong (1913, as cited in Valencia & Suzuki, 2001, Chapter 1) reported White-Negro differences on intelligence. Although this difference was interpreted by some as evidence of the superiority of the White race (see discussion below), others interpreted this mean score difference as evidence that the tests were biased. Problems with the a priori assumptions associated with this interpretation are addressed in depth by Jensen (1980) in his discussion of the "egalitarian fallacy"; however, the issue of test bias was not settled. A number of psychologists, particularly Black and minority psychologists, raised objections to the use of intellectual tests with minorities (e.g., Thomas, 1982; Valencia, 1997). Reynolds and Lowe (2009, p. 339-340) and Reynolds, Lowe, & Saenz (1999, p. 556-557) identified seven categories into which the most frequently stated problems fall; these categories and a summary of their descriptions are included in Table 6 (see Reynolds & Lowe, 2009 and Reynolds, Lowe, & Saenz, 1999 for more details).

Table 6

Categories and Descriptions of Most Frequently Stated Problems with the Use of Intelligence Tests with Ethnic Minorities

Category	Description
Inappropriate content	Tests and test materials are rooted in mainstream, majority culture and values. Racial and ethnic minorities in the US may not have been exposed to content in test questions or stimulus materials, which may result in cognitively equivalent responses that are scored as incorrect. Thus, scores may reflect differences in values, not ability.
Inappropriate standardization samples	Historically, it was not unusual for standardization samples to be all White, and thus racial/ethnic minorities were underrepresented in these samples. Contemporary standardization samples are more often representative based on race/ethnicity proportions in the national population.
Examiners' and language bias	Most psychologists in the United States are White and speak only standard English, which may mean they are unable to accurately communicate with minority children—"to the point of being intimidating and insensitive to ethnic pronunciation of words on the test" (p. 339).
Inequitable social consequences	A long history of discrimination and being thought unable to learn (i.e., limited opportunity to learn) in conjunction with bias in educational and psychological testing may lead to ethnic and racial minorities being disproportionately represented in dead-end/low achievement educational tracks.
Measurement of different constructs	Similar to the points made in the first category, educational and psychological tests may measure different attributes for ethnic/racial minority children as compared to the majority culture, which negatively impacts the validity of test results.
Differential predictive validity	Measures may not accurately predict outcomes for minority group members, despite appearing to accurately predict outcomes for majority group members. Further criticisms that the way outcome criteria are defined (i.e., outcome criteria are inherently biased) are also included in this category.
Qualitatively distinct minority and majority aptitude and personality	Ethnic minority cultures and the majority culture may be so different as to require different conceptualizations of ability and personality, and thus would require separate tests for different groups.

Note. Table provides summary of the categories and descriptions provided by Reynolds & Lowe (2009, p. 339-340) and Reynolds, Lowe, & Saenz (1999, p. 556-557).

One weakness identified in early objections was that they were "frequently stated as facts on rational rather than empirical grounds" (e.g., Reynolds & Brown, 1984, p. 17). However, in contrast to when the controversy began, empirical research has been conducted to investigate the validity of these claims. Reviews of this literature can be found elsewhere (e.g. Jensen, 1980; Reynolds, 1982b), though some of this evidence will be incorporated into the subsequent section. In general, the categories identified by Reynolds and Lowe (2009) remain relevant and capture the essence of the problems identified over time with regard to test bias, particularly within the domain of cultural bias.

Historical precedent. Although a detailed discussion regarding definitions of test bias will be presented in a later section, it is important to note at this point that in general, bias in intelligence testing has historically referred to cultural bias (both racially oriented and class oriented, see e.g., Valencia & Suzuki, 2001, Chapter 5), which will also be the case throughout this review of the historical precedent of bias in intelligence testing. This section will acknowledge selected key ideological influences and historical developments that may have contributed to bias in intelligence testing, as well as describe specific historical events in which test bias was addressed overtly. The content within this "Historical Precedent" section was obtained from numerous sources, however, the organization, including section headings, is similar to that used by Valencia and Suzuki (2001, Chapter 1).

Ideology of the intelligence testing movement. The fundamental contributions of measurement to the intelligence testing movement are commonly recognized (see, e.g., Cohen and Swerdlik, 2002; Sattler, 2001). The impact of ideology on intelligence testing, however, does not receive as much attention, despite arguments that it may have contributed to shaping the intelligence testing movement to an equal degree as measurement (see, e.g., Cravens, 1978; Degler, 1991; Fass, 1980; Gould, 1981; Guthrie, 1976; Marks, 1981; as cited in Valencia, 1997 and Valencia & Suzuki, 2001, Chapter 1). As the individual credited with initiating the psychological testing movement, Galton is a natural candidate for consideration when investigating potential sources of ideological influence. As described previously in this review, much of Galton's work in psychology was inspired by Darwin's writings on natural selection, and his hereditarian perspective is common knowledge. One observation that has not yet been made in this review, however, was the overt presence of racist views presented by Galton. In his book *Hereditary Genius* (1869), for example, Galton included a chapter entitled, "The Comparative Worth of Different Races," in which he provided a list of races ranked according to his observations. He concluded that ancient Greeks were considered the "ablest race of whom history bears record" (p. 340). Anglo-Saxons were slightly lower, with the "African negro" below the lowest Anglo-Saxon, and the "Australian type" "one grade below the African negro" (p.339). Despite including only two relatively short chapters focused explicitly on race, a "scientific racist" perspective is clearly apparent (Richards, 1997).

Valencia and Suzuki (2001, Chapter 1) identified two primary influences of Galton's biased views on the field of psychological testing, "(a) how American psychologists approached the practice of testing minority children on intelligence tests (largely a practice of indifference to cultural and environmental differences), and (b) how behavioral scientists and applied psychologists attempted to explain White-minority group differences in intellectual performance (frequent genetic interpretations)" (p. 4-5). Goddard, Terman, and Yerkes, all hereditarians (and all mentioned previously in this review), are prime examples of the infiltration of Galton's views into the discourse of psychological and intelligence testing (Gould, 1996). Terman's comments in his guide for the clinical application of the original Stanford-Binet (*Measurement of Intelligence,* 1916) provide a very clear illustration of this influence. His statements below follow a description of two Portuguese brothers included for the purposes of illustrating "borderline" intelligence (individuals with IQs between 70 and 80):

> It is interesting to note that M.P. and C.P. represent the level of intelligence that is very, very common among Spanish-Indian and Mexican families of the Southwest and also among negroes. Their dullness seems to be racial, or at least inherent in the family stocks from which they come. The fact that one meets this type with such extraordinary frequency among Indians, Mexicans, and negroes suggests quite forcibly that the whole question of racial differences in mental traits will have to be taken up anew and by experimental methods. The writer predicts that when this is done, there will be discovered

> enormously significant differences in general intelligence, differences that cannot be wiped out by any scheme of mental culture.
>
> Children of this group should be segregated in special classes and be given instruction that is concrete and practical. They cannot master abstractions, but they often can be made efficient workers, able to look out for themselves.
>
> There is no possibility at present of convincing society that they should not be allowed to reproduce, although from a eugenic point of view they constitute a grave problem because of their unusually prolific breeding (p. 91-92, also as cited in Valencia & Suzuki, 2001, Chapter 1).

The presence of racial bias in Terman's statements is unequivocal. Moreover, the presence of Galton's ideological influence is undeniable as well. Although specific implications of these views can be somewhat difficult to identify, one possible secondary consequence of the popularity of these views is potentially, in part, the facilitation of the reification of the concept of intelligence. Although Gould (1981) attributes this reification primarily to Terman's creation of the "Intelligence Quotient," without the prominent hereditarian views of the time, the tendency toward viewing intelligence as a concrete (and completely measurable) construct may not have been so powerful.

Because the biggest forces contributing to the momentum of the testing movement were (generally speaking) hereditarians, eugenicists, and often racists, "it makes sense to conclude that intelligence testing—consisting of procedural, measurable, and social movement aspects—was a value-laden idea with significant implication for the stratification of schooling practices and outcomes" (Valencia & Suzuki, 2001, Chapter 1,

p. 7). This evidence suggests that there has not only been at least some degree of ideological influence on the intelligence testing movement since its inception, but, more specifically, there has also been a culturally biased ideology serving as a significant contributor to the foundation of intelligence testing. For a more detailed discussion of this ideological influence and its implications, see e.g., Valencia's (2010) discussion of neohereditarianism.

An extension of these early ideological influences can be attributed to the "race psychology" studies conducted during the initial stages of the testing movement. "Race psychology," a term coined by Thomas R. Garth in his research on this movement, refers to psychological research conducted with the objective of making comparisons among different races. This field can be considered a precursor to what is presently called cross-cultural research, though one distinction is that many of these early studies were conducted by researchers openly ascribing to hereditarian values and frequently providing hereditarian explanations for any differences identified between groups (e.g., Richards, 1997; Valencia, 2010). Garth conducted two comprehensive reviews of the race psychology literature, one in 1925 reviewing research publications spanning 1916 to 1924, and another in 1930 reviewing publications from 1924 to 1929 (Garth, 1925, 1930, as cited in Valencia & Suzuki, 2001, Chapter 1). Although there were several domains examined in the studies Garth included in his reviews, one consistent trend across both reviews was a primary area of focus in the literature on making comparisons between White children and children of color with regard to their intellectual performance. As Valencia and Suzuki (2001, Chapter 1) explain, in Garth's earlier review of the literature

(1925), he concluded that, when considered together, the studies suggested "mental superiority of the white race" (p. 359), later termed the "hypothesis of racial inequality" (Garth, 1930, p. 348). However, Garth's later review of the literature suggested that the popularity of the "hypothesis of racial inequality" appeared to be deteriorating while a "hypothesis of racial equality" was garnering more support (1930). As Valencia & Suzuki, 2001, Chapter 1 indicate, the majority of the studies in Garth's 1925 review offered hereditarian explanations for group differences, while none of the studies in Garth's 1930 review did, which suggests that there was in fact, a shift away from hereditarianism, or "scientific racism" (Richards, 1997).

Heterodoxy. Although hereditarian views were the predominant paradigm of the time, it is important to acknowledge that there were also researchers who deviated from the status quo (Valencia, 1997; Valencia, 2010). As mentioned previously, Binet presented warnings regarding the limitations of his measures, though his warnings were largely ignored, as evidenced by the mass (mis)use of his measure. One individual who was able to have a more significant impact in terms of actively opposing hereditarian views was a prominent researcher by the name of Otto Klineberg (Richards, 1997), who worked in the 1920s and 1930s to "answer, or at least to offer responses, to practically all existing research that proffered the position that certain groups were inferior" (Valencia, 1997b, p. 17). In his book, *Race Differences* (1935), Klineberg offered critiques of racial superiority theories and emphasized the need for considerations of culture in explaining and understanding group differences. Additionally, Klineberg offered one of the early methodological critiques of testing research at the time, identifying and supporting

empirically geographical location, English-language abilities, and socioeconomic status as confounds in existing research at the time. Klineberg thus concluded that any results suggesting superiority of one group over another (namely White groups over minority groups) were invalid, as an acceptable degree of methodological rigor was lacking.

Thomas (1982) describes additional dissenting work, in this case contributed in the 1920s by Black researchers who offered strong critiques of mental testing at the time. Thomas organizes their work into three categories: environmental critiques, identification of methodological flaws or instrumentation weaknesses, and the implementation of independent research and the generation of additional data. As an aside, these categories are somewhat similar to those of Reynolds and Brown (2009) presented earlier, although Thomas's categories directly incorporate empirical evidence. Environmental critiques encompass environmental characteristics as explanations for differences between racial groups: for example, differences in educational opportunity between Whites and Blacks can account for racial differences in test performance (e.g., Bond, 1924, as cited in Thomas, 1982). Technical criticisms of intelligence tests were quite extensive, as indicated, for example, by Long (1923, as cited in Thomas, 1982), who argued that tests of intelligence contain numerous measurement problems. Finally, independent research was able to investigate alternative explanations to hereditarian conclusions, eg. Bond (1924, as cited in Thomas, 1982), who investigated schooling effects on mental test performance, and Canady (1936, as cited in Thomas, 1982), who was one of the first researchers to explore examiner effects on test performance across racial groups.

As Valencia (2001, Chapter 1) noted, there were no Mexican American scholars who actively criticized mental testing research until 1931, when George Isodore Sánchez entered the field. Sánchez's (1934) essay is identified by Valencia (2001, Chapter 1) as being one of his most significant articles, in which Sánchez provided a critique of mental measures and testing as they relate to Mexican American students. Padilla and Aranda (1974, p. 222, as cited in, Valencia and Suzuki, 2001, Chapter 1, p.21-22) summarize the seven key issues raised by Sánchez:

1) Tests are not standardized on the Spanish-speaking population of this country.
2) Test items are not representative of the Spanish-speaking culture.
3) The entire nature of intelligence still is a controversial issue.
4) Test results from the Spanish-speaking continue to be accepted uncritically.
5) Revised or translated tests are not necessarily an improvement on test measures.
6) Attitudes and prejudices often determine the use of test results.
7) The influence of testing on the educational system is phenomenal. (p. 21)

Emergence of contemporary testing issues. In general, the period between 1930 and the mid-1950s was characterized by maintenance of the status quo regarding test bias. There was a gradual decline in the acceptance of hereditarianism/scientific racism in scientific domains, though the underpinnings remained present in some ways (Richards, 1997; Valencia, 2010). Efforts of dissenters like those mentioned above continued as well, and the second phase of test bias controversy did not occur until 1954, when the US Supreme Court released

its decision for the *Brown v. Board of Education of Topeka* (Valencia & Suzuki, 2001, Chapter 1). This decision struck down the former "separate but equal" doctrine of *Plessy v. Ferguson* (1896) but in itself did not directly affect the test bias movement. Rather, the response of Southern schools to this ruling was ultimately what set the ball in motion. In their reluctance to integrate their schools, Southerners used the administration of intelligence and achievement measures to preclude entry of Black children into "white" schools (Bersoff, 1982, as cited in Valencia & Suzuki, 2001, Chapter 1). Through legislation these activities were eventually determined to be unconstitutional. Interestingly in this case, only the issue of test *use* was under fire; the validity of the tests themselves was never brought into question (Bersoff, 1982, as cited in Valencia & Suzuki, 2001, Chapter 1).

With the civil rights movement highlighting the rights of racial/ethnic minorities at approximately the same time as this attempt to use intelligence measures to circumvent desegregation, there was something of a "perfect storm" that helped to rekindle the test bias debate (e.g., Valencia and Suzuki, 2001, Chapter 1). Two primary features of this debate were the relationships between (a) curriculum differentiation (i.e., tracking) and group-administered intelligence tests, and (b) overrepresentation of minority students in special education and individually-administered intelligence tests (Valencia, 1999; Valencia, 2008).

Hobson v. Hanson (1967), the first case focusing on the legality of using group-administered intelligence tests as the sole criteria for determining curriculum placement, marks the beginning of a series of cases claiming that intelligence tests were being used

in a racially/ethnically discriminatory fashion. Valencia and Suzuki (2001, Chapter 5; see also, Valencia, 2008) highlight four of these cases, (a) *Diana v. State Board of Education* (1970), (b) *Covarrubias v. San Diego Unified School District* (1971), (c) *Guadalupe v. Tempe Elementary School District* (1972), and (d) *Larry P. v. Riles* (1972, 1979) as particularly significant due to their roles in bringing attention to the issue of overrepresentation of minority students in educable mentally retarded classes (Henderson & Valencia, 1985; Valencia, 1999; Valencia, 2008). These four victorious cases, brought forth by African American, Mexican American, and American Indian plaintiffs, all contained arguments primarily addressing the validity of the tests themselves, as the plaintiffs asserted certain items were biased, rather than the use or application of the tests. Especially interesting is that these "analyses for bias" were based primarily on the subjective opinion of one or more individuals regarding whether the item at "face value" appeared to be biased. As Valencia and Suzuki (2001, Chapter 5) observe, at that point in time, scientific investigation into test bias had not yet begun.

In response to criticisms similar to those outlined in an earlier section (Reynolds & Lowe, 2009) regarding potential problems in the appropriateness of using intelligence measures on minorities, in conjunction with advances in statistical methods for evaluating test bias (see e.g., Millsap & Meredith, 2007), the mid-1970s through the 1980s were characterized by a surge of empirical investigations of test bias. There has been a slight decline in test bias research since the beginning of the 1990s, perhaps due to conclusions presented by Jensen (1980) suggesting that after an "exhaustive review of the empirical research bearing on this issue…the currently most widely used standardized tests of

mental ability…are, by and large, *not* biased against any of the native-born English-speaking minority groups" (p. ix). This decline may also be associated with increased attention on issues of bias during measurement development and revision, as preliminary investigations of bias are being conducted prior to test publication. Nonetheless, the importance of this research has not waned, and investigations into test bias continue. A review of existing test bias literature will be presented later.

Bias: What it is and what it is not. The primary focus within this review up to this point has been relatively broad, reviewing issues related to bias from several perspectives, and generally melding views coming from a variety of definitions. Within the literature, the best methods of defining and evaluating bias has been a source of persistent controversy and has been largely ignored in this review up to this point. Thus, the objective of the following section is to address similar or related constructs that are not, technically speaking, "bias," to discuss various definitions of test bias in the literature, and to clarify its definition for the purposes of the proposed study.

Cultural loading vs. cultural bias. As Reynolds and Brown (1984) note, the concepts of cultural loading and cultural bias are often (incorrectly) used interchangeably, even within the professional literature. Cultural loading, they state, "refers to the degree of cultural specificity present in the test or in individual items of the test" (p. 23). Although one can assume that a greater degree of cultural loading (or a greater level of cultural specificity) of a test item will increase the likelihood that the item is biased toward individuals from outside of that culture, cultural loading does not necessitate the presence of cultural bias. Cultural bias as a result of cultural loading, as implied by this

example, refers to a mismatch between the cultural foundations of the test and the cultural background or cultural exposure of the individual taking the test. Thus, cultural bias does not refer to the mere presence of culture loading within a measure, although attempts have been made to reduce the degree of culture loading on various measures in order to minimize the potential for culture bias.

One issue with attempting to reduce cultural loading on intelligence tests is the problem of measuring cultural loading in the first place. Jensen (1980) suggests that cultural loading is often not addressed because subjective judgments ("arm chair analyses") by individuals are the primary method for identifying these influences. However, an additional perspective that makes the evaluation of cultural loading seem somewhat less urgent is that all mental tests in use are culturally bound to some degree and may inevitably be so due to psychometric standardization methods (e.g., Harrington, 1975; 1976; 1984). Although this is sometimes viewed as a major flaw or weakness of intellectual assessment measures, the opposite, a "culture blind" test, cannot "be expected to predict intelligent behavior within a cultural setting" (Reynolds & Brown, 1984, p. 23). Thus, given that all intelligence measures are necessarily developed within a culture, the generalizability of those measures to other cultures or subcultures must be investigated empirically.

Primary selection models and test fairness. Academic debates within the literature have produced a number of models through which test bias can be understood (eg. Cleary, Humphreys, Kendrick, & Wesman, 1975; Darlington, 1971; Reynolds & Brown, 1984; Thorndike, 1971; Van de Vijver and Leung, 1997a,b; Van de Vijver and

Poortinga, 1997; Van de Vijver and Tanzer, 2004). One such model is the primary selection model, which as a whole tends to focus on the decision-making system associated with assessment, rather than the test itself. The choice of which decision-making system to employ is largely a societal issue, as it is a heavily value- and goal-driven issue (e.g., Reynolds & Brown, 1984; Reynolds & Lowe, 2009). Nichols (1978) elaborates on this idea, suggesting that in order to choose a model for use in selection, the ultimate goal—whether for equality of opportunity, equality of outcome, or representative equality— must first be established. Despite the merits of discussing these issues, these questions reach beyond the scope of this review.

The question of test fairness is similar to that regarding primary selection models, in that the question that arises is one that goes beyond the scope of the measure itself; test fairness, like various selection models, is a function of the application or interpretation of a test, as opposed to a function of the test itself. Some argue that this perspective supports a clear distinction between test fairness (a subjective value judgment) and test bias (conceptualized as being objective and technical), making test fairness and its opposite, test unfairness, "belong more to moral philosophy than to psychometrics" (Jensen, 1980, p. 49). Jensen comments further:

> Unbiased tests can be used unfairly and biased tests can be used fairly. Therefore, the concepts of bias and unfairness should be kept distinct… [A] number of different, and often mutually contradictory, criteria of fairness have been proposed, and no amount of statistical or psychometric reasoning per se can possibly settle any arguments as to which is best. (p. 375-376)

Although Jensen (1980) conceptualizes these issues as completely separate, others argue that test bias and test fairness are linked to varying degrees (e.g., Cole & Moss, 1989; Hilliard, 1984; Mercer, 1984). In either case, one necessary prerequisite for appropriate application and interpretation of intelligence tests (both in the interest of test fairness and an accepted selection model) is to employ measures that possess equal psychometric reliability and validity for all groups concerned (Reynolds & Brown, 1984; Reynolds & Lowe, 2009).

Test bias: Defined. Based on the preceding discussion of selection models, test fairness, and cultural loading, three primary conclusions can be drawn: 1) because intelligence tests are inherently culture-bound, generalizability of these measures to other cultures and subcultures must be investigated empirically, 2) test bias must refer to characteristics of the test itself, rather than the application or interpretation of the measure, and 3) measures of intelligence must possess equal psychometric reliability and validity for all (relevant) groups (Brown, Reynolds, & Whitaker, 1999). In considering these conclusions as criteria for defining test bias, a definition of test bias naturally ensues: " systematic error in the estimation of some "true" value for a group of individuals" (Reynolds & Lowe, 2009).

This definition of test bias— "the systematic (not random) error of some true value of test scores that are connected to group membership" (Valencia & Suzuki, 2001, Chapter 5, p. 115)—is consistent with the psychometric view of test bias and can be investigated empirically. As a psychometric definition by nature, this explanation of test bias is a statistical definition "that does not concern itself with culture loading, labeling

effects, or test use or fairness" (Reynolds & Lowe, 2009, p. 345), exists independently of any test, and, like many definitions of test bias, is closely linked to test validity. It is important to recognize that there are other possible definitions for test bias (as mentioned at the beginning of the "Primary Selection Models and Test Fairness" section of this review; see also Reynolds and Lowe, 2009 for a more detailed discussion/review); however, none of these alternatives meets all of the aforementioned criteria, and thus the above definition of test bias will be assumed for the remainder of this review and was used for the purposes of the current study.

Relationship between test bias and validity. As noted above and corroborated by Cole and Moss (1989), "many definitions of bias [e.g., Reynolds, 1982b, Shepard, 1981, 1982] have been closely tied to validity theory" (p. 205); moreover, empirical research investigating test bias often applies a paradigm of validity, either a traditional "tripartite conceptualization"—of content, construct, and predictive or criterion-related validity—or a "dyadic conceptualization"—of internal and external validity (e.g., Messick, 1980, 1981; Reynolds & Brown, 1984; Van de Vijver, & Tanzer, 2004). This conceptualization of test validity as a "unitary" construct has achieved support by a number of measurement experts (e.g., Cole & Moss, 1989; Cronbach, 1980; Messick, 1981, 1989), though there are some limitations to the tripartite model.

One such limitation is the problem of oversimplification (e.g., Messick, 1980), as it often leads to the incorrect assumption that establishing one type of validity indicates that a test is "valid" (or unbiased) overall. In short, no one type of validity is sufficient to establish a lack of test bias. In light of this, Reynolds and Brown (1984) suggested that a

scientific model of validity would "almost certainly have as one of its requirements the hierarchical arrangement of the subsets of validity" (p. 22), with content validity established before construct validity, and construct validity established before predictive validity. That being said, there is some agreement that construct validity is the unifying concept of test validity (Reynolds & Brown, 1984), as "all validity is at its base some form of construct validity" (Guion, 1977, p. 410) due to its ability to integrate "criterion and content considerations into a common framework for testing rational hypotheses about theoretically relevant relationships" (Messick, 1980, p. 1015).

Although validity and test bias are obviously connected, an important distinction to make between the two is that bias involves comparison between two (or more) groups, while validity can apply to only one group (Jensen, 1980). This observation is especially relevant when considering the "test bias" definitions of each type of validity (in a subsequent section), as empirical test bias can apply to any type of group (i.e. groups based on race/ethnicity, sex, socioeconomic status, etc.). Generally, however, when one is comparing ethnic or racial groups it is referred to as cultural bias (eg. Valencia & Suzuki, 2001, Chapter 5). Furthermore, when comparing ethnic or racial groups, the comparison is typically between two groups, often with one identified as the "major" group and the other labeled as the "minor" group. Jensen (1980) clarifies these distinctions, as they are not indicative of value judgments:

> The major group can usually be thought of as (1) the larger of the two groups in the total population, (2) the group on which the test was primarily standardized, or

(3) the group with the higher mean score on the test, assuming the major and minor groups differ in means (p. 376).

Valencia and Suzuki (2001, Chapter 5) add an additional distinction: "The major group is the group that the test is believed not to be biased against" (p. 117).

The next sections will address (to a limited extent) the manner in which these comparisons are made by presenting "test bias" definitions of the specified types of validity, followed by a brief description of some of the more popular methods used to test for the specified type of bias. Some strengths and weaknesses of each method are also included. For the purposes of organization, the tripartite conceptualization of validity will be used, as this method is consistent with other reviews (e.g., Reynolds & Lowe, 2009). More detailed reviews can be found elsewhere (e.g., Camilli & Shepard, 1994; Jensen, 1980; Reynolds, 1982b, 1995, 2000; Reynolds & Brown, 1984).

Content bias. Test bias in terms of content validity has been empirically defined with a testable definition by Reynolds (1982a):

> An item or subscale of a test is considered to be biased in content when it is demonstrated to be relatively more difficult for members of one group than another when the general ability level of the groups being compared in held constant and no reasonable theoretical rationale exists to explain group differences on the item (or subscale) in question. (p. 188)

Common methods employed to identify cultural bias at the item level include the following:

Analysis of variance (ANOVA), a popular method for identifying cultural bias—the most popular, in fact, until the late 1980s (Camilli & Shepard, 1987)—is used to evaluate whether a significant group by item interaction exists, which would suggest that item difficulty is not uniform across groups (Reynolds & Lowe, 2009). Camilli and Shepard (1994) evaluated this method, and based on significant limitations, concluded that it should not be used to detect cultural bias at the item level. *Item response theory (IRT),* intended to determine differential item functioning across groups, was the foundation of the replacement methodology recommended by Camilli and Shepard (1994) after thorough analysis and their rejection of ANOVA methods. The *partial correlation* method, developed independently by Stricker (1982) and Reynolds, Wilson, and Chatman (1985), assesses group differences by determining the amount of variation in the observed scores not due to the total score. Although not as popular as IRT, partial correlation procedures have been used frequently (Jensen, 1980) and have been shown to effectively detect cultural bias at the item level (Reynolds, Lowe, & Saenz, 1999). Finally, *expert approaches,* in which expert reviewers assess items for content bias, are often used during test development (e.g., Kaufman & Kaufman, 2004a, 2004b; Elliott, 2007). Significant weaknesses in this method have been identified, as empirical studies did not support the ability of expert reviewers to predict differential item functioning (Camilli & Shepard, 1994; Reynolds, 2000).

Construct bias. Reynolds (1982a) provided the following definition of construct bias in terms of construct validity:

> Bias exists in regard to construct validity when a test is shown to measure different hypothetical traits (psychological constructs) for one group than another or to measure the same trait but with different degrees of accuracy. (p. 194)

By nature, the evaluation of measures for construct bias requires research from a variety of viewpoints with a range of methodologies (Reynolds, 1982a). Thus, several methods have been employed to examine tests for potential construct bias.

Factor analytic methods qualify as one of the most popular types of empirical approaches to investigating construct bias (Anastasi, 1996; Cronbach, 1990; Millsap, 2011). These methods allow investigators to determine patterns of interrelationships of performance across groups. In general, exploratory factor analysis methods are typically more useful in the application of tests to diagnosis, while confirmatory factor analysis techniques tend to be employed more in the case of hypothesis-testing research. Although there is some debate regarding whether these methods capture the innateness of the abilities measured, there is general agreement that consistent results across groups provide robust evidence that the construct measured by the test is a) measured in the same way for both groups, and b) actually the same construct for each group (Reynolds & Lowe, 2009).

Reliability methods are another group of techniques that can be used to assess construct bias. These methods include internal consistency reliability, test-retest reliability, and alternate forms reliability (e.g., Kane, 2006). Internal consistency reliability estimates— representing "the degree to which the items on a scale or subscale are all measuring a similar construct" (Reynolds & Lowe, 2009, p. 356)— should be

approximately equal across groups to demonstrate a lack of bias in construct validity. Like internal consistency estimates, test-retest correlations across groups should be similar to indicate a lack of bias in construct validity. Differences between groups on test-retest correlations does not necessarily indicate bias, however, as these differences can also be a result of practice effects or instability in the trait measured. Alternate forms reliability is primarily used in less common cases in which coefficient alpha or Kuder-Richardson 20 (frequently used measures of internal reliability) are not appropriate or not possible to compute (Reynolds & Lowe, 2009).

Predictive bias. The third and final component of the tripartite conceptualization of validity, predictive bias, is defined by Cleary, Humphreys, Kendrick, & Wesman (1975):

> A test is considered biased with respect to predictive validity when the inference drawn from the test score is not made with the smallest feasible random error or if there is constant error in an inference or prediction as a function of membership in a particular group. (p. 201)

This definition of predictive bias is considered a "regression definition" and is thus best assessed using *simple regression*. Bias in prediction using regression is determined to be absent when the regression equation is the same across groups. Significant differences in the slope or intercept across groups will result in biased prediction if a regression equation for the combined groups is used. Several methods are also available for situations in which multiple predictors and/or multiple criterion variables are involved (Reynolds & Lowe, 2009). *Path analysis* was proposed by Keith and Reynolds (1990) as

an alternative method for evaluating bias in predictive validity. Via this method, consideration of the degree to which errors of measurement in testing of ability correlate with group membership is used to determine whether there is evidence of predictive bias. It should be noted, however, that although predictive bias is above content and construct in the hierarchical framework, an absence of predictive bias does not necessarily imply that there is an absence of measurement bias (Wicherts & Millsap, 2009)

Existing test bias research. Several authors have conducted extensive reviews of the test bias literature (see, e.g., Jensen, 1980; Reynolds & Lowe, 2009; Valencia & Suzuki, 2001, Chapter 5); however, the most recent, systematic, and detailed review was that conducted by Valencia and Suzuki (2001, Chapter 5). This review is intended to build on theirs. Thus, a brief summary of their findings is provided, followed by an updated review of research conducted since the publication of their work.

Using frequency data collected from each of the studies reviewed, Valencia and Suzuki were able to make several astute and significant observations regarding the cultural bias research to date. One overarching pattern they identified among the studies reviewed is that identification of cultural bias appeared to be "psychometric specific." More precisely, studies investigating predictive bias tended to find more evidence of cultural bias, whereas those studies investigating construct bias (and reliability) tended to find very little evidence of cultural bias. These findings indicate the potential for confounds relating to the methodologies for investigating the various types of validity bias, or perhaps relating to the definitions used for conceptualizing bias. Via their review, Valencia and Suzuki identified a number of weaknesses in cultural bias research.

These limitations included, (a) a lack of research investigating the potential for bias in minority groups other than Black and Hispanic, (b) limited geographical representation of the United States (only 12 of the 50 states were represented), (c) potential problems with external validity, due to the primary representation within the research of children in general education, while most of the children being tested with the measures in question have been referred for special education, and (d) failure to control for socioeconomic status, language dominance, and/or sex in most studies.

In an effort to provide an updated review of the literature the procedures reported by Valencia and Suzuki were employed, with the result being that only five studies conducted since 1999 met their criteria for inclusion (as compared to the 62 studies identified in their 2001 review). The review of these studies follows a format similar to that of Valencia and Suzuki, again, for the sake of consistency and ease of comparison. Table 7 includes descriptive data about these studies that is similar to the types of data provided in Valencia and Suzuki's original review.

Table 7

Characteristics of Cultural Bias Research Since 1999

#	Authors	Year	Type of Validity	Measure	Minor Group Participants	Finding
1	Keith, Quirk, Schartzer, & Elliott	1999	Construct	Differential Ability Scales	Black; Hispanic	Non-bias
2	Floyd, Gatheroal, & Roid	2004	Content	Merill-Palmer Scale, Tryout Edition	Black; Hispanic	Non-bias
3	Edwards & Oakland	2006	Construct	Woodcock-Johnson-III	Black	Non-bias
4	Qi & Marley	2009	Content	Preschool Language Scale-4	Hispanic	Non-bias
5	Konold & Canivez	2010	Predictive	Wechsler Intelligence Scale for Children-IV; Wechsler Individual Achievement Test-II	Black; Hispanic	Non-bias

Results: Study characteristics. *Year of study.* In consideration of the year of study, one can observe that there tends to be a slightly heavier loading of studies (3) during the second half of the 2000s than in the earlier 2000s (1) or late 1990s (1). This observation may be an optimistic indicator of renewed interest in conducting investigations of cultural bias in intelligence measures. A second possibility is that the trend coincides with the development and release of new editions of tests, in which case there should be some acknowledgment of the associated test bias research; however, it is doubtful that these five studies sufficiently examine all new and updated versions of cognitive measures published since 1999.

Psychometric property examined. The distribution of psychometric properties studied is somewhat skewed, as content and construct validity were both explored in two studies each, whereas predictive validity was investigated once, and assessing bias through reliability methods was not employed at all. The lack of reliability research in

the cultural bias domain may possibly be accounted for by the limited conclusions regarding test bias that can be drawn through this type of investigation. An alternative explanation may also be that test reliability is frequently a focus during test development, and many published measures may already have established sound reliability across groups.

Intelligence test examined. Three of the five studies explored bias in fairly well-known, batteries for school-age children (and older, in some cases), including the Wechsler Intelligence Scale for Children-IV, the Woodcock-Johnson-III Test of Cognitive Abilities, and the Differential Ability Scales. The other two measures are designed for younger children, with the Preschool Language Scale-4 used for children 3-5 years old and the Merrill-Palmer Scale for children ages 18 months to 4 years.

Race/Ethnicity of minor group. All five studies were limited to comparisons with Black, Hispanic, or both with regard to the race/ethnicity of the minor group. This trend is consistent with both expectations and limitations recognized by Valencia and Suzuki (2001, Chapter 5), as these two groups are the largest minority groups in the United States, but other minority populations (e.g., Native Americans and Asian Americans) continue to be underrepresented in test bias research.

Number of participants. The total number of participants across all five studies appears to have been 5,737 children. Of these, 3,278 (57.1%) were major group children and 2,459 (43.0%) were minor group children (1165- 20.3%-Hispanic; 1294- 22.5%-Black). Similar to the pattern observed by Valencia and Suzuki (2001, Chapter 5), there was considerable variability in sample size across studies, with the smallest study

including 295 total participants and the largest having almost 2,400 total participants. Overall, it appears that sample sizes used for the analyses appear to have been of sufficient size to provide adequate power for the associated statistical analyses.

Remaining characteristics. Additional characteristics discussed by Valencia and Suzuki included gender, age, socioeconomic status (SES), language status, and education status. Three of the 5 studies reported data on sex for the total sample, with one study providing additional data regarding sex across major and minor groups. Four of the 5 studies reported the age of participants in some form, with the fifth study reporting grade level (K-12). The range of ages across these studies is quite broad, spanning from 34 months to 17 years, 11 months, and only two studies conducted any form of comparison across age groups. Regarding SES, two studies reported having "representative" samples, but neither specifically mentioned SES as one of the criteria used for determining "representative". One study reported a sample comprised entirely of "low-income" participants, and though the specific criteria for this label was not disclosed, all participants were enrolled in a Head Start program, which are designed for economically disadvantaged populations. Two studies reported categorical breakdowns of SES based on parent education level. Only one study reported education status of participants, as the authors clarified that children participating the study did not have individual education plans outside of their speech and language impairments. This same study was the only investigation conducted with data collected outside of a standardization/norming sample and the only study that reported language dominance of children (monolingual English).

Results: Test bias. *Content bias*. Studies 2 and 4 conducted investigations into content bias, with both investigations using differential item functioning methods to identify items with potential bias. Both studies identified 2-4 questionable items, though the authors concluded that their investigations did not identify large-scale (significant) systematic item bias against either the minor or major group.

Specifically, the authors of Study 2 identified two items (of 320 total items) as being differentially more difficult for one group and with a large effect size. One item (for 3-year-olds) suggested a longer completion time by White children as compared to Black children, and the other item (for 4-year-olds) which more Hispanic children were able to correctly complete as compared to White children. It should be noted that the measure used for Study 2 was a tryout edition, and at the time of publication of the article, it was unknown whether the items identified would remain on the final published version of the measure.

Study 4 used a MANOVA to compare group means on subscales and the total scale initially, which indicated that their sample performed statistically lower than the national norms and that there were no mean differences between groups on any subscale or total scale across White and Hispanic groups. Item level analyses identified four items with differential functioning, with two items identified as more difficult for Hispanic children, one item more difficult for White children, and one item with non-uniform differential functioning. The authors conducted an "arm-chair" content analysis of these items but were unable to identify an explanation and thus concluded that although these items may show differential functioning, they are not necessarily biased.

Construct bias. Studies 1 and 3 investigated construct bias. Study 1 conducted multi-group confirmatory factor analysis to assess measurement invariance; whereas study 3 conducted principal factor analysis, multiple-group structural equation modeling, r_c analysis, and convergent validity analysis. Conclusions for both studies were generally in support of a lack of bias across groups; however, it appears that Study 1 reported stronger results in support of a lack of bias.

Specifically, Study 1 reported strong support for the construct validity of the measure under scrutiny (the Differential Ability Scales: DAS); however, the findings regarding construct bias showed some variability across ethnic groups depending on age. For youth ages 2 to 3 ½ and ages 12 to 17, there appeared to be no evidence of construct bias, but for children ages 3 ½ to 5 and 6 to 11, results suggested that the test may measure different constructs across ethnic groups. It should be noted, however, that the authors investigated these apparent differences further and suggest that the differences identified across groups are due to variability in the unique and error variances of the subtests, which are expected to vary across groups, rather than differences in the constructs measured across groups. Thus, the authors suggest that the evidence supports the conclusion of a lack of construct bias on the measure.

Study 3 suggested that the measure under consideration (the Woodcock-Johnson, 3rd edition: WJ-III) has a factor structure that is consistent across Black and White groups and does not display evidence of bias across these two groups. Limitations of Study 3 included the use of only one score for each of the seven cognitive abilities in the analysis (using two scores would likely improve the reliability of the results), and the analyzing of

correlations to determine the probability that differences in correlations were significant (as this method does not take into account possible differences in variable and factor variances). Thus, the authors of Study 3 concluded that factorial invariance was established, but in this case does not necessarily preclude the presence of group differences in performance on the measure that may affect test interpretation.

Predictive bias. Study 5 investigated criterion-related bias (or differential prediction bias) of scores across race, gender, and parent education level. Concurrent criterion-related correlations between the measures were large and statistically significant across all groups. Moderately large to large coefficients were observed when the total sample was considered and across subgroups according to race/ethnicity, gender, and parent education. Differential relationships were observed for five of the 30 total comparisons, but none were observed across race/ethnic groups or gender for the WIAT-II and WISC-IV FSIQ scores. Effect sizes for statistically significant differential relationships were all small, and the authors indicated that the few differences observed are of limited clinical importance. One strength of this study is that the authors investigated additional variables beyond major and minor group comparisons.

Although the distribution of bias and non-bias findings do not appear to be "psychometric specific" like they were in Valencia and Suzuki's review (2001, Chapter 5), the small number of studies precludes the proposal of any potential alternative conclusions. These five studies were generally consistent with Valencia and Suzuki's (2001, Chapter 5) findings, however, with regard to the paucity of rigorous research characterizing more recent decades. As Valencia and Suzuki observe, research

investigating test bias appears to be on the decline, particularly when one compares the number of studies conducted in the 1970s and 1980s to the number conducted in the 1990s forward. This observation is consistent with the current state of the literature, as the 1970s produced 24 studies, the 1980s produced 33, the 1990s produced 6, and the 2000s produced only 4. Although it seems unlikely based on the recent trend, perhaps the 2010s will experience a resurgence of culture bias research. Despite the recent decline in the popularity of this type of research, the importance to "press on" has not waned (Suzuki & Valencia, 1997).

Valencia and Suzuki identified cultural bias investigations for fourteen different intelligence measures, many of which have since been revised once (e.g. Elliott, 2007; Kaufman & Kaufman, 2004) or in some cases twice (e.g., Wechsler, 1991, 2003; Woodcock & Johnson, 1989; Woodcock, McGrew, & Mather, 2001). Most of these updated and current measures have not been evaluated for the presence of cultural bias. Given that intelligence tests are (a) among the most popular measures that psychologists administer (Stinnett, Havey, & Oehler-Stinnett, 1994; Wilson & Reschly, 1996); (b) used for purposes as diverse as determining eligibility for special education services, identification of an individual's need for services, determination of parameters (i.e. intensity and duration) of treatment (Dowdy, Mays, Kamphaus, & Reynolds, 2009), political advocacy, program evaluation, and research (Keough, 1994); and (c) likely to be administered with increasing frequency to diverse populations as globalization redefines boundaries of the world and the demographics within the United States evolve, the importance of evaluating measures for cultural bias cannot be emphasized enough.

References

Anastasi, A. (1996). *Psychological testing* (7th ed.). New York: Macmillian.

Arbuckle, J. L. (2006). Amos (Version 19.0) [Computer Program]. Chicago: SPSS.

Archer, R.P., Maruish, M., Imhof, E.A., & Piotrowski, C. (1991). Psychological test usage with adolescent clients: 1990 survey findings. *Professional Psychology: Research and Practice, 22,* 247-252.

Bayley, N. (2005). *Bayley Scales of Infant and Toddler Development.* San Antonio, TX: The Psychological Corporation.

Binet, A., & Simon, T. (1905). Methods nouvelle pour le diagnostic de nivea intellectual des anormaux. *L'Ann Ce Psychologique, 11,* 191-244.

Binet, A., & Simon, T. (1916/1973). *The development of intelligence in children.* New York: Arno.

Blum, J.M. (1978). *Pseudoscience and mental ability: The origins and fallacies of the IQ controversy.* New York: Monthly Review Press.

Boake, C. (2002). From the Binet-Simon to the Wechsler-Bellevue: Tracing the history of intelligence testing. *Journal of Clinical and Experimental Neuropsychology, 24,* 383-405.

Bond, L. (1987). The golden rule settlement: A minority perspective. *Educational Measurement Issues and Practice, 6,* 23-25.

Boomsma, A. (2000) Reporting analyses of covariance structures. *Structural Equation Modeling, 7,* 461-483.

Boring, E.G. (1923, June). Intelligence as the tests test it. *New Republic,* pp. 35-37.

Brannick, M. T. (1995). Critical comments on applying covariance structure modeling. *Journal of Organizational Behavior, 16,* 201-213.

Brown v. Board of Education of Topeka, 347 U.S. 483 at 494 (1954).

Brown, R.T., Reynolds, C.R., & Whitaker, J.S. (1999). Bias in mental testing since *Bias in Mental Testing. School Psychology Quarterly, 14,* 208-238.

Browne, M., & Cudeck, R. (1993). Alternative ways of assessing model fit. *Sociological*

Methods and Research, 21, 230–258.

Buros, O.K. (Ed.). (1941). *The 1940 mental measurements yearbook.* Highland Park, NJ: Gryphon Press.

Buros, O.K. (1949). *The third mental measurements yearbook.* Highland Park, NJ: Gryphon Press.

Butler-Omololu, C., Doster, J., & Lahey, B. (1984). Some implications for intelligence test construction and administration with children of different racial groups. *Journal of Black Psychology, 10,* 63-75.

Camara, W.J., Nathan, J.S., & Puente, A.E. (2000). Psychological test usage: Implications in professional psychology. *Professional Psychology: Research and Practice, 31,* 141-154.

Camilli, G., & Shepard, L.A. (1987). The inadequacy of ANOVA for detecting test bias. *Journal of Educational Statistics, 12,* 87-99.

Camilli, G., & Shepard, L.A. (1994). *Methods for identifying biased test items.* Thousand Oaks, CA: Sage.

Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Cattell, R.B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin, 38,* 592.

Chapman, P.D. (1988). *Schools as sorters: Lewis M. Terman, applied psychology, and the intelligence testing movement, 1890-1930.* New York: New York University Press.

Cheung, G.W., & Rensvold, R.B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9,* 233-255.

Cleary, T.A., Humphreys, L.G., Kendrick, S.A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 30,* 15-41.

Cohen, R.J., & Swerdlik, M.E. (2002). *Psychological testing and assessment: An introduction to tests and measurement* (5th ed.). Mountain View, CA: Mayfield Publishing Company.

Cole, N.S., & Moss, P.A. (1989). Bias in test use. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201-219). New York: Macmillan.

Covarrubias v. San Diego Unified School District, Civil Action No. 70-30d (S.D. Cal.) (1971).

Cronbach, L.J. (1980). Validity on parole: How can we go straight? In W.B. Schrader (Ed.) *New directions for testing and measurement,* No. 5, *Measuring achievement: Progress over a decade* (pp. 99-108). San Francisco: Jossey-Bass.

Cronbach, L.J. (1990). *Essentials of psychological testing* (5th ed.). New York: Harper & Row.

Darlington, R.B. (1971). Another look at "cultural fairness." *Journal of Educational Measurement, 8,* 71-82.

Davidson, J.E. & Downing, C.L. (2000). Contemporary models of intelligence. In R.J. Sternberg (Ed.), *Handbook of intelligence* (pp. 34-49). New York: Cambridge University Press.

Diana v. State Board of Education, Civil Action No. C-70-37 (N.D. Cal.) (1970).

Dowdy, E., Mays, K.L., Kamphaus, R.W., & Reynolds, C.R. (2009). Roles of diagnosis and classification in school psychology. In C.R. Reynolds & T.B. Gutkin (Eds.), *The handbook of school psychology* (pp. 191-209). New York: Wiley.

Edwards, O.W. & Oakland, T.D. (2006). Factorial invariance of Woodcock-Johnson III scores for African Americans and Caucasian Americans. *Journal of Psychoeducational Assessment, 24,* 358-366.

Eells, R.L., Davis, A., Havighurst, R.J., Herrick, V.E., & Taylor, R.W. (1951). *Intelligence and cultural differences: A study of cultural learning and problem-solving.* Chicago: University of Chicago Press.

Elliott, C.D. (1990). *Differential Ability Scales: Introductory and technical manual.* San Antonio, TX: The Psychological Corporation.

Elliott, C.D. (1996). *British Ability Scales, Second edition.* Windsor, England: NFER-Nelson.

Elliott, C.D. (2007). *Differential Ability Scales, Second edition.* San Antonio, TX: The Psychological Corporation.

Elliott, C.D., Murray, D.J., & Pearson, L.S. (1979). *British Ability Scales.* Windsor, England: National Foundation for Educational Research.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6,* 56-83.

Floyd, R.L., Gathercoal, K., & Roid, G. (2004). No evidence for ethnic and racial bias in the tryout edition of the Merrill-Palmer Scale-Revised. *Psychological Reports, 94,* 217-220.

Galton, F. (1869/1892/1962). *Hereditary genius: An inquiry into its laws and consequences*. London: Macmillan and Company.

Galton, F. (1883/1907/1973). *Inquiries into human faculty and its development*. New York: AMS Press.

Goddard, H.H. (1910). A measuring scale of intelligence. *Training School, 6,* 146-155.

Goddard, H.H. (1917). Mental tests and the immigrant. *Journal of Delinquency, 2,* 243-277.

Gould, S.J. (1981/1996). *The mismeasure of man.* New York: Norton.

Graham, J.W., & Coffman, D.L. (2012). Structural equation modeling with missing data. In R.H. Hoyle (Ed.), *Handbook of structural equation modeling* (pp. 277-295). New York: Guilford Press.

Guadalupe v. Tempe Elementary School District, No. 3, Civ. No. 71-435 (D. Ariz.) (1972).

Guilford, J.P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Guion, R.M. (1977). Content validity: Three years of talk—Where's the action? *Public Personnel Management, 6,* 407-414.

Gustaffson, J.E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8,* 179-203.

Harrington, G. M. (1975). Intelligence tests may favour the majority groups in a population. *Nature, 258,* 708-709.

Harrington, G. M. (September, 1976). *Minority test bias as a psychometric artifact: The experimental evidence.* Paper presented to the annual meeting of the American Psychological Association, Washington, D.C.

Harrington, G. M. (1984). *The new American poverty.* New York: Holt, Rinehart, and Winston.

Helms, J.E., Jernigan, M., & Mascher, J. (2005). The meaning of race in psychology and how to change it: A methodological perspective. *American Psychologist, 60,* 27-36.

Henderson, R.W., & Valencia, R.R. (1985). Nondiscriminatory school psychological services: Beyond nonbiased assessment. In J.R. Bergan (Ed.), *School psychology in contemporary society* (pp. 340-377). Columbus, OH: Merrill.

Hilliard, A.G., III. (1984). IQ testing as the emperor's new clothes: A critique of Jensen's *Bias in Mental Testing.* In C.R. Reynolds & R.T. Brown (Eds), *Perspectives on bias in testing* (pp. 139-169). New York: Plenum.

Hobson v. Hanson, 269 F. Supp. 401 (D.C. 1967) aff'd sub. Nom., Smuck v. Hobson, 408 F.2d 175 (D.C. Cir.) (1969).

Horn, J.L., & Cattell, R.B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57,* 253-270.

Horn, J.L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D.P. Flanagan, J.L. Genshaft, & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 53-91). New York: Guilford Press.

Hu, L., & Bentler, P.M. (1998). Fit indexes in covariance structure modeling: Sensitivity to underparameterized model misspecifcation. *Psychological Methods, 3,* 424-453.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6,* 1-55.

IBM Corp. Released 2010. IBM SPSS Statistics for Windows, Version 19.0 [Computer Program]. Armonk, NY: IBM Corp.

Jensen, A.R. (1980). *Bias in mental testing.* New York: Free Press.

Jensen, A.R. (1998). *The g factor: The science of mental ability.* Wesport, CT: Praeger.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17-64). Library of Congress Cataloging-in-Publication Data.

Kaufman, A.S., & Kaufman, N.L. (2004a). *Kaufman Assessment Battery for Children* (2nd ed.). Circle Pines, MN: American Guidance Service.

Kaufman, A.S., & Kaufman, N.L. (2004b). *Kaufman Test of Educational Achievement* (2nd ed.). Circle Pines, MN: American Guidance Service.

Keith, T.Z. (2006). *Multiple regression and beyond.* New York: Pearson.

Keith, T.Z., Fugate, M.H., DeGraff, M., Diamond, C.M., Shadrach, E.A., & Stevens, M.L. (1995). Using multi-sample confirmatory factor analysis to test for construct bias: An example using the K-ABC. *Journal of Psychoeducational Assessment, 13,* 347-364.

Keith, T.Z., Low, J.A., Reynolds, M.R., Patel, P.G., & Ridley, K.P. (2010). Higher-order factor structure of the Differential Ability Scales-II: Consistency across ages 4 to 17. *Psychology in the Schools, 47,* 676-697. doi:10.1002/pits.20498.

Keith, T.Z., Quirk, K.J., Schartzer, C., & Elliott, C.D. (1999). Construct bias in the Differential Ability Scales: Confirmatory and hierarchical factor structure across three ethnic groups. *Journal of Psychoeducational Assessment, 17,* 249-268.

Keith, T.Z., & Reynolds, C.R. (1990). Measurement and design issues in child assessment research. In C.R. Reynolds & R.W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children.* New York: Guilford.

Keith, T.Z., & Reynolds, M.R. (2010). Cattell-Horn-Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools, 0,* 1-16.

Keith, T. Z., & Reynolds, M. R. (2012). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 758-799). New York: Guilford.

Keith, T.Z., Reynolds, M.R., Roberts, L.G., Winter, A.L., & Austin, C.A. (2011). Sex differences in latent cognitive abilities ages 5 to 17: Evidence from the Differential Ability Scales—Second Edition. *Intelligence, 39,* 389-404.

Kelloway, E. K. (1995). Structural equation modeling in perspective. *Journal of Organizational Behavior, 16*, 215-224.

Keough, B.K. (1994). A matrix of decision points in the measurement of learning disabilities. In G.R. Lyon (Ed.), *Frames of reference for the assessment of learning disabilities: New views on measurement issues* (pp. 15-26). Baltimore, MD: Paul H. Brooks.

Kline, R.B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York: Guilford.

Klineberg, O. (1935). *Race differences.* New York: Harper.

Konold, T. R., & Canivez, G.L. (2010). Differential relationships between WISC-IV and WIAT-II Scales: An evaluation of potentially moderating child demographics. *Education and Psychological Measurement, 70,* 613-627.

Larry P. v. Riles, 343 F Supp. 1306 (N.D. Cal. 1972, order granting preliminary injunction), aff'dd 502 F.2d 63 (9th Cir. 1974), 495 F. Supp. 926 (N.D. Cal. 1979, decision on merits), aff'd No. 80-427 (9th Cir. Jan.23, 1984), No. C-71-2270 R.F.P. (Sept. 25, 1986, order modifying judgment).

Lubin, B., Larsen, R.M., & Matarazzo, J.D. (1984). Patterns of psychological test usage in the United States: 1935-1982. *American Psychologist, 39,* 451-454.

Lubin, B., Wallis, R.R., & Paine, C. (1971). Patterns of psychological test usage in the United States: 1935-1969. *Professional Psychology: Research and Practice, 2,* 70-74.

McDonald, R.P., & Ho, M.-H.R. (2002). Principles and practice in reporting statistical equation analyses. *Psychological Methods, 7 (1),* 64-82.

McGrew, K.S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities: Past, present, and future. In D.P. Flanagan & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136-182). New York: The Guilford Press.

McIntosh, D.E., & Dixon, F.A. (2005). Use of intelligence tests in the identification of giftedness. In D.P. Flanagan & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 504-520). New York: The Guilford Press.

Mercer, J.R. (1979). In defense of racially and culturally non-discriminatory assessment. *School Psychology Digest, 8,* 89-115.

Mercer, J.R. (1984). What is a racially and culturally nondiscriminatory test? A sociological and pluralistic perspective. In C.R. Reynolds & R.T. Brown (Eds.), *Perspectives on bias in mental testing* (pp. 293-356). New York: Plenum.

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58,* 525-543.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012-1017.

Messick, S. (1981). Constructs and their vicissitudes in educational and psychological measurement. *Psychological Bulletin, 89,* 575-588.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Millsap, R.E. (2011). *Statistical approaches to measurement invariance.* New York, NY: Routledge.

Nichols, R.C. (1978) Policy implications of the IQ controversy. In L.S. Schulman (Ed.), *Review of research in education* (Vol. 6). Itsaca, IL: F.E. Peacock.

Pintner, R. (1931). *Intelligence testing: Methods and results* (2nd ed.). New York: Henry Holt.

Plessy v. Ferguson, 163 U.S. 537 (1896).

Qi, C.H., & Marley, S.C. (2009). Differential item functioning analysis of the Preschool Language Scale-4 between English-speaking Hispanic and European American children from low-income families. *Topics in Early Childhood Special Education, 29,* 171-180.

Reynolds, C. R. (1982a). Methods for detecting construct and predictive bias. In R.A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 199-297). Baltimore, MD: Johns Hopkins University Press.

Reynolds, C.R. (1982b). The problem of bias in psychological assessment. In C.R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 178-208). New York: Wiley.

Reynolds, C.R. (1995). Test bias in the assessment of intelligence and personality. In D. Daklofske & M. Zeidner (Eds.), *International handbook of personality and intelligence* (pp. 545-576). New York: Plenum.

Reynolds, C.R. (2000). Methods for detecting bias in neuropsychological tests. In E. Fletcher-Janzen, T.L. Strickland, & C.R. Reynolds (Eds.), *Handbook of cross-cultural neuropsychology* (pp. 249-285). New York: Kluwer Academic.

Reynolds, C.R., & Brown, R.T. (1984). Bias in mental testing: An introduction to the issues. In C.R. Reynolds & R.T. Brown (Eds.), *Perspectives on Bias in Mental Testing* (pp. 1-39). New York: Plenum.

Reynolds, C.R., & Lowe, P.A. (2009). The problem of bias in psychological assessment. In C.R. Reynolds & T.B. Gutkin (Eds.), *The handbook of school psychology,* (4th ed., pp. 332-374). New York: Wiley.

Reynolds, C.R., Lowe, P.A., & Saenz, A. (1999). The problem of bias in psychological assessment. In C.R. Reynolds & T.B. Gutkin (Eds.), *The handbook of school psychology,* (3rd ed., pp. 549-595). New York: Wiley.

Reynolds, C.R., Wilson, V.L., & Chatman, S.P. (1985). Regression analyses of bias on the Kaufman Assessment Battery for Children. *Journal of School Psychology, 23,* 195-204.

Richards, G. (1997). *Race, racism, and psychology: Towards a reflexive history.* New York, NY: Routledge.

Sattler, J.M. (2001). *Assessment of children: Cognitive applications.* San Diego, CA: Jerome M. Sattler, Publisher.

Siegler, R. S. (1992). The other Alfred Binet. *Developmental Psychology, 28,* 179-190.

Spearman, C. (1904). General intelligence objectively determined and measured. *American Journal of Psychology, 15*, 201–293.

Steiger, J.H. (1998). A note on multiple sample extensions of the RMSEA fit index. *Structural Equation Modeling, 5,* 411-419.

Stern, W. (1914). *The psychological methods of testing intelligence.* Baltimore, MD: Warwick & York.

Stinnett, T.A., Havey, J.M., & Oehler-Stinnett, J. (1994). Current test usage by practicing school psychologists: A national survey. *Journal of Psychoeducational Assessment, 12,* 331-350.

Stricker, L.J. (1982). Identifying test items that perform differentially in population subgroups: A partial correlation index. *Applied Psychological Measurement, 6,* 261-273.

Suzuki, L.A., & Valencia, R.R. (1997). Race-ethnicity and measured intelligence: Educational implications. *American Psychologist, 52,* 1103-1114.

Terman, L.M. (1916). *The measurement of intelligence: An explanation of and a complete guide for the use of the Stanford Revision and Extension of the Binet-Simon Intelligence Scale.* Cambridge, MA: Houghton.

Terman, L.M., & Merrill, M.A. (1937). *Measuring intelligence.* Boston: Houghton Mifflin.

Terman, L.M., & Merrill, M.A. (1960). *Stanford-Binet Intelligence Scale: 1960 norms edition.* Boston: Houghton Mifflin.

Thomas, W.B. (1982). Black intellectuals' critique of early mental testing: A little known saga of the 1920s. *American Journal of Education, 90,* 258-292.

Thorndike, R.L. (1971). Concepts of culture-fairness. *Journal of Educational Measurement, 8,* 63-70.

Thurstone, L.L. (1938). The absolute zero in intelligence measurement. *Psychological Review, 35,* 175-197.

Thurstone, L.L. (1947). Multiple factor analysis. Chicago: University of Chicago Press.

Tuddenham, R.D. (1962). The nature and measurement of intelligence. In L.J. Postman (Ed.), *Psychology in the making (pp. 469-525). New York: Knopf.*

Tulsky, D. (2003). Reviews and promotional material for the Wechsler-Bellevue and Wechsler Memory Scale. In D.S. Tulsky, et. al. (Eds.), *Clinical interpretation of the WAIS-II and WMS-III* (pp. 579-602). San Diego, CA: Academic Press.

Valencia, R.R. (1997). Genetic pathology model of deficit thinking. In R.R. Valencia (Ed.), *The evolution of deficit thinking: Educational thought and practice* (pp. 41-112). Stanford Series on Education and Public Policy. London: Falmer.

Valencia, R.R. (1999). Educational testing and Mexican American students: Problems and prospects. In J.F. Moreno (Ed.), *The elusive quest for equality: 150 years of Chicano/Chicana education* (pp. 123-139). Cambridge, MA: Harvard Educational Review.

Valencia, R.R. (2008). *Chicano Students and the Courts: The Mexican American Legal Struggle for Educational Equality.* New York, NY: New York University Press.

Valencia, R.R. (2010). *Dismantling contemporary deficit thinking*. New York, NY: Routledge.

Valencia, R.R. (2013). Jason Richwine's dissertation, IQ and immigration policy: Neohereditarianism, pseudoscience, and deficit thinking. *Teachers College Record*. Retrieved from http://www.tcrecord.org

Valencia, R.R., & Aburto, A. (1991). The uses and abuses of educational testing: Chicanos as a case in point. In R.R. Valencia (Ed.), *Chicano school failure and success: Research and policy agendas for the 1990s* (pp. 203-251). Stanford Series on Education and Public Policy. London: Falmer.

Valencia, R.R., Rankin, R.J., & Livingston, R. (1995). K-ABC content bias: Comparisons between Mexican American and White children. *Psychology in the Schools, 32,* 153-169.

Valencia, R.R., & Suzuki, L.A. (2001). *Intelligence testing and minority students: Foundations, performance factors, and assessment issues*. Thousand Oaks, CA: Sage.

Van de Vijver, F.J.R., & Leung, K. (1997). Methods and data analysis of comparative research, In: J.W. Berry, Y.H. Poortinga, & J. Pandey (Eds.), *Handbook of cross-cultural psychology* (pp. 257–300). Boston: Allyn & Bacon.

Van de Vijver, F.J.R., & Leung, K., (1997). *Methods and data analysis for cross-cultural research.* Newbury Park, CA: Sage.

Van de Vijver, F.J.R., & Poortinga, Y.H., (1997). Towards an integrated analysis of bias in cross-cultural assessment. *European Journal of Psychological Assessment 13,* 29–37.

Van de Vijver, F., & Tanzer, N.K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *European Review of Applied Psychology, 54,* 119-135.

Vernon, P.E. (1961). *The measurement of abilities* (2nd ed.). Oxford: Philosophical Library.

Wasserman, J.D., & Tulsky, D.S. (2005). A history of intelligence assessment. In D.P. Flanagan & P.L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 1-38). New York: The Guilford Press.

Wechsler, D. (1939a). *The measurement of adult intelligence.* Baltimore: Williams & Wilkins.

Wechsler, D. (1939b). *Wechsler-Bellevue Intelligence Scale.* New York: The Psychological Corporation.

Wechsler, D. (1949). *Manual for the Wechsler Intelligence Scale for Children.* New York: The Psychological Corporation.

Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale.* New York: The Psychological Corporation.

Wechsler, D. (1967). *Manual for the Wechsler Preschool and Primary Scale of Intelligence.* New York: The Psychological Corporation.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children—Third Edition.* New York: The Psychological Corporation.

Wechsler, D. (2001). *Manual for the Wechsler Individual Achievement Test—Second Edition.* New York: The Psychological Corporation.

Wechsler, D. (2003). *Manual for the Wechsler Intelligence Scale for Children—Fourth Edition.* San Antonio, TX: The Psychological Corporation.

Weiss, L. G., Keith, T. Z., Zhu, J., & Chen, H. (2013). WAIS-IV and clinical validation of the four- and five-factor interpretive approaches. *Journal of Psychoeducational Assessment, 31*, 94-113.

Wicherts, J.M., & Millsap, R.E. (2009). The absence of underprediction does not imply the absence of measurement bias. *American Psychologist, 64,* 281-283.

Wilson, M.S., & Reschly, D.J. (1996). Assessment in school psychology training and practice. *School Psychology Review, 25,* 9-23.

Woodcock, R.W., & Johnson, M.B. (1989). *WJ-R Tests of Cognitive Abilities.* Itasca, IL: Riverside.

Woodcock, R.W., McGrew, K.S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Cognitive Abilities.* Itasca, IL: Riverside.

Yerkes, R.M. (1921). *Psychological examining in the United States Army* (Memoirs of the National Academy of Sciences, Vol. 15). Washington, DC: Government Printing Office.

Yoakum, C.S., & Yerkes, R.M. (1920). *Army mental tests.* New York: Holt.